19 November 2004

Science

Vol. 306 No. 5700
Pages 1241–1420 $10

Fundamentals of MEASUREMENT

AAAS

Science

Special Issue

Fundamentals of Measurement

Volume 306
19 November 2004
Number 5700

Metrology, the science of measurement, provides the basis for the set of standard metrics by which we can make accurate measurements. A special section in this issue of *Science* explores advances in metrology, the development of the tools required, and the ways in which these tools may eventually affect our lives. [Photo: JupiterImages]

Departments

News of the Week

News Focus

1278

1304

Letters

Books *et al.*

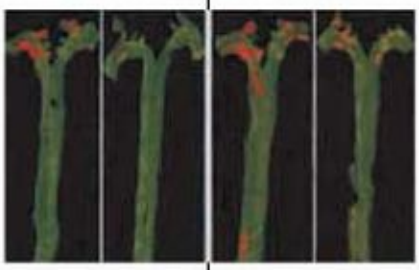
Policy Forum

Perspectives

Contents continued

Perspectives continued

Science Express www.sciencexpress.org



1273 &1339

Technical Comment Abstracts

Brevia

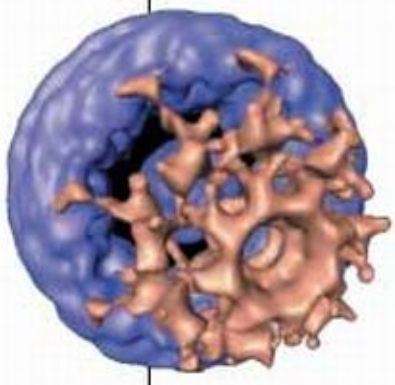
Research Articles

Reports

1275 &1362

Contents continued

REPORTS CONTINUED



1387

1390

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals Mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2004 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership and subscription (51 issues): $130 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $550. Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery) $85. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #1254 88122. Publications Mail Agreement Number 1069624. Printed in the U.S.A.

Change of address: allow 4 weeks, giving old and new addresses and 8-digit account number. Postmaster: Send change of address to Science, P.O. Box 1811, Danbury, CT 06813–1811. Single copy sales: $10.00 per issue prepaid includes surface postage; bulk rates on request. Authorization to photocopy material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that $15.00 per article is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923. The identification code for Science is 0036-8075/85 $15.00. Science is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

Contents continued

Science
www.scienceonline.org

sciencenow www.sciencenow.org Daily News Coverage

Watch Your Step
New measurements reveal how water-running lizards avoid tripping.

Less Fish, More Meat
Smaller fish harvests have helped drive the bushmeat trade in Ghana.

Second Black Hole for the Milky Way
Midsize black hole could explain puzzling stars near the galaxy's core.

Helping science students understand business.

science's next wave www.nextwave.org Career Resources for Young Scientists

US: Tooling Up—Doing/Thinking/Judging *D. Jensen*
Earl Wagoner's mission is to help students make a successful transfer from academia to industry.

MiSciNet: Tale of Two Chemists—Finding Fulfillment in Science *E. Francisco*
Two research chemists explain how their love of math and science led them to exciting careers in consumer products.

GrantsNet: November 2004 Funding News *Edited by S. Martin*
Learn about funding opportunities including the Pasteur Foundation's postdoc fellowships in Paris.

Canada: Canadian Science Bytes *A. Fazekas*
Read about the latest funding, training, and job market news from Canada.

Europe: European Science Bytes *Next Wave Staff*
Read about the latest funding, training, and job market news from Europe.

US: Careers in Science Web Log *J. Austin*
Breaking news and observations related to science careers are updated throughout the week.

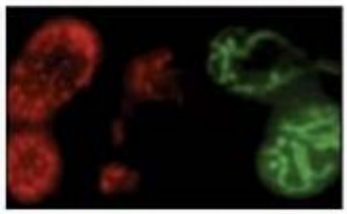
science's sage ke www.sageke.org Science of Aging Knowledge Environment

Experimental Rodent Strain: F344BNF1 and BNF344F1 Hybrid Rats *D. J. Holmes*
Hybrid strains are longer-lived and healthier than the parental strains.

News Synthesis: Culture Clash *R. J. Davenport*
Yeast cells might sign suicide pact, but the idea remains controversial.

News Focus: Ironing Out Cell Death *M. Leslie*
Metal-grabbing enzyme saves suicidal cells.

News Focus: Solving the Energy Crisis *R. J. Davenport*
Signaling molecule spurs cells to generate mitochondria.



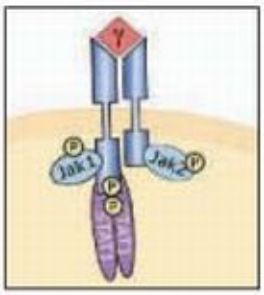
Do yeast rise to the art of dying?

science's stke www.stke.org Signal Transduction Knowledge Environment

Teaching Resource: The Jak-STAT Pathway Stimulated by Interferon γ *C. M. Horvath*
This animation shows stimulation of gene expression by a type II interferon.

Teaching Resource: The Jak-STAT Pathway Stimulated by Interleukin 6 *C. M. Horvath*
This animation shows stimulation of gene expression by the inflammatory cytokine IL-6.

Teaching Resource: The Jak-STAT Pathway Stimulated by Interferon α or Interferon β *C. M. Horvath*
This animation shows stimulation of gene expression by the antiviral type I interferons.



Signaling by type II interferons.

Separate individual or institutional subscriptions to these products may be required for full-text access.

Toward Optical Metamaterials

Metamaterials that are engineered to exhibit negative indices of refraction can provide a number of advantages in optics, such as the fabrication of a "perfect" lens, and much effort is being directed to push the frequencies at which negative indices can be achieved into the optical regime. Using nanofabrication techniques to shrink the dimensions of gold nanostructures making up the metamaterial, **Linden *et al.*** (p. 1351) show that the magnetic response can be raised to 100 terahertz. On the theoretical side, **Pendry** (p. 1353) introduces an alternate route to the design of metamaterials exhibiting negative refraction that may prove easier to prepare than the present structures, which are based on tuning the electric and magnetic response. The proposed structure relies on chirality and consists of a series of helically folded metallic foils. Designers should be able to work with the polarization of either the magnetic or the electric field, rather than both.

Improving Optical Clocks

The development of frequency-stabilized sources of laser radiation, together with the associated coupling of frequency cycles in the optical regime, offer the potential to exceed the accuracy set by atomic standards that operate in the lower frequency microwave regime. **Margolis *et al.*** (p. 1355) have developed an optical frequency standard based on measuring the transition frequency of a trapped strontium ion. The transition frequency is determined to nearly 1 Hertz in 10^{15} and represents a fractional uncertainty within a factor of three of the primary cesium atomic-clock standards.

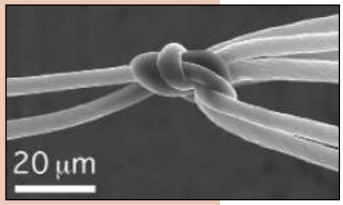
The Spin on Martian Argon

Argon concentration in the martian atmosphere can be used to trace the planet's rotational dynamics and seasonal patterns. Measurements from the Gamma-Ray Spectrometer on the Mars Odyssey spacecraft by **Sprague *et al.*** (p. 1364, published online 7 October 2004; see the Perspective by **Forget**) suggest that non-condensible argon is enhanced at mid-latitudes during the summer and decreases at more polar latitudes in early autumn, even though carbon dioxide is condensing out of the atmosphere onto the polar cap in the southern hemisphere. The data suggest that meridional mixing occurs, which is contrary to the idea that separate vortices of material, particularly at the poles, are driven by planetary rotation.

CREDITS: (TOP TO BOTTOM) ZHANG *ET AL.*; FORGET

Nanotube Yarns and Forests

Spinning fibers to make stronger yarns and ropes is an old technology. By looking at the fundamentals of this process and scaling them down to fibers with nanometer-sized diameters, **Zhang *et al.*** (p. 1358) have developed a technique to spin carbon nanotube yarns from mats of fibers. The twisted yarns can be infiltrated with a polymer to improve their strength. Unlike larger diameter materials, a knot can be made in the ropes without a loss of strength. Many methods have been developed for making single-walled carbon nanotubes, but there are still significant limitations to making the tubes in large quantities and free from impurities or residual catalysts. **Hata *et al.*** (p. 1362; see the news story by **Service**) modified the standard chemical vapor deposition synthesis by adding a small amount of water, which removes the residual carbon from the catalyst particles and keeps them chemically active for longer periods. The nanotube forests are easily removed from the bed of catalyst particles, which continue to be chemically active.



Before the Divide with the Great Apes

The group that includes humans and great apes is thought to have diverged from other apes (such as gibbons) in the Middle Miocene, about 10 to 15 million years ago. Few relatively complete fossils are available from this time; all are thought to be related to later great apes from Eurasia. **Moyà-Solà *et al.*** (p. 1339; see the news story by **Culotta**) have recovered a remarkably preserved fossil of a new ape species from Spain dating to about 13 million years ago. The cranium, which is nearly complete and undistorted, the thorax, and bones including the wrist show a mix of both primitive, derived, and very modern features. The skeleton also shows that the distinctive posture of great apes had evolved by this time. The fossil may be close to the last common ancestor of the great apes and humans.

Understanding Mimi

Mimivirus is an extremely large DNA virus that grows in amoebae. **Raoult *et al.*** (p. 1344, published online 14 October 2004) have sequenced and analyzed the genome of the Mimivirus, which is 1.2 megabases long—more than three times larger than any other viral genome previously sequenced. Among its 1200 open reading frames are genes not previously thought to be part of the classical definition of a viral repertoire, including genes with homology to transfer RNAs (tRNAs), translation initiation factors, polysaccharide synthesis enzymes, tRNA synthetases, and enzymes involved in nucleic acid metabolism. Mimivirus appears to represent a new family of nucleocytoplasmic large DNA viruses that emerged early in evolution.

Identifying the Chosen Strand

Small interfering (si)RNAs provide the sequence information that allows the RNA-induced silencing complex (RISC) to destroy target messenger RNAs. siRNAs generated by the enzyme Dicer are double-stranded (ds), but the "guide" RNA used by RISC needs to be a single stand. The stability of the base pairs at the 5′ ends of both of the siRNA strands plays an important role in distinguishing between them. **Tomari *et al.*** (p. 1377) now provide insight into how this choice is made. The RISC loading complex, which consists of Dicer itself together with the dsRNA-binding protein, R2D2, can detect and

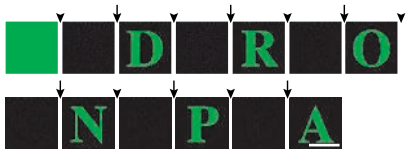
CONTINUED ON PAGE 1257

Published by AAAS

differentiate between the siRNA 5′ end stabilities, with R2D2 binding to end with the most double-stranded character. As the siRNA is unwound, the guide strand would then be transferred from R2D2 to RISC, while the other strand would be destroyed.

Now You See It, Now You Don't

Existing fluorescent protein highlighting techniques are irreversible and preclude repeated monitoring of the same protein to study its temporal regulation. Within cells, protein movement is regulated by many different factors and may be altered by changes in the cellular state. Measurements of protein dynamics are affected by the geometry of both the cells and the highlighted regions, and any changes in movement should ideally be assessed using data from a single cell. **Ando *et al.*** (p. 1370) describe the engineering and application of a fluorescent protein, Dronpa, which can be reversibly highlighted to study spatiotemporal protein dynamics in living cells. The authors directly visualized the influx and efflux of a key regulator of intracellular signaling, mitogen-activated protein kinase, into and out of the nucleus.



Virus Exploits a Serotonin Receptor

JC virus (JCV) is a common human polyomavirus responsible for the fatal demyelinating disease, progressive multifocal leukoencephalopathy (PML), in immunocompromised individuals—about 5% of AIDS patients develop this currently untreatable fatal disease. Typical and atypical antipsychotic drugs inhibit JCV infection of glial cells. **Elphick *et al.*** (p. 1380) now find that the cellular receptor for JCV on glial cells is a serotonin receptor. The findings contribute to the understanding of the pathogenesis of PML in AIDS patients and suggest that therapy based on existing serotonin receptor inhibitors may be feasible.

Please Release Me

During times of food deprivation or increased energy demand, mammals begin to use the intracellular triglycerides stored in fat tissue as a primary energy source. Mobilization of these stored lipids requires activation of enzymes that degrade them so that free fatty acids, the molecules that supply energy to most tissues, are released into the blood. **Zimmermann *et al.*** (p. 1383) identify a new enzyme, adipose triglyceride lipase (ATGL), that is expressed at high levels in mammalian adipose tissue and catalyzes the initial step in triglyceride degradation. Because abnormalities in lipid metabolism are often associated with obesity and type 2 diabetes, ATGL could represent an important new drug target for these conditions.

The Nuclear Pore, Up Close and Personal

Cryoelectron tomography of intact cells or organelles has been developed to study molecular structures in their native environments, unaffected by isolation and purification procedures which may entail the loss of components. **Beck *et al.*** (p. 1387, published online 28 October 2004) studied intact nuclei from *Dictyostelium discoideum* by cryoelectron tomography with a focus on the structure of the nuclear pore complexes. The images detail the components of the pore and reveal putative transport substrates.

Regulating Oscillatory Calcium Signals

Variation in the intensity and frequency of intracellular calcium signals impact numerous calcium-dependent cellular responses, but the underlying mechanisms that regulate oscillatory calcium signaling have not been fully resolved. **Launay *et al.*** (p. 1374) report that generation and maintenance of the calcium oscillations that control the production of the cytokine interleukin-2 in stimulated T cells involve a calcium-activated nonselective cation channel called TRPM4. In response to a rise in intracellular calcium, TRPM4 is activated and contributes to depolarization of the membrane potential, which suppresses further calcium influx. Subsequent repolarization closes TRPM4 channels and reestablishes conditions for further calcium influx.

CREDIT: ANDO *ET AL.*

Published by AAAS

Science in the South

The International Centre for Theoretical Physics (ICTP) in Trieste, Italy, turned 40 this October. It is an occasion for some reflection. The scientists who created ICTP, notably the Nobel Laureate Abdus Salam of Pakistan, were motivated by a goal that is simple to proclaim but difficult to fulfill: to advance the level and role of science in the Southern world by overcoming the debilitating isolation of scientists who work there.

This goal is more important now than ever before. No country today can survive and prosper in isolation, and economic prosperity is tied to scientific development. The building of scientific capacity needed everywhere is thus in our collective interest and is a shared responsibility. Forty years on, however, we still live in a world in which a majority of scientists, scientific discoveries, publications, and patents come from developed countries. So, what has ICTP accomplished?

ICTP has been involved, to different degrees, with the careers of some 100,000 visiting scientists. They have come from nearly every country in the world, about half from developing countries. According to physics professor Edmund Zingu of Mangosuthu Technikon in South Africa, "Nearly every Ph.D. in East Africa has had an association with ICTP." The cadre of ICTP associates has established programs in their home countries, including Brazil, Benin, China, India, and Mexico. Some have turned to public service as ministers of science, members of parliaments, ambassadors, and in one case, the president of a republic. ICTP thus exemplifies that the best investment one can make is in human capital: the individual scientists.

Building scientific capacity worldwide is a shared responsibility.

But ICTP is keenly aware that its efforts are small relative to the needs. These needs are tremendous even in countries that have made some strides (at least progress has been spotty). Regrettably, countries in Africa and the Middle East have either stood still in scientific progress or actually regressed. The challenges remain daunting. The critical question is how to proceed.

We can draw one lesson: Among the diverse ways in which ICTP has attempted to fulfill its mission, the key ingredient for success has been the followthrough. Where we have been able to keep sustained contact with our associates, the success has been greater. Because ICTP is small, large-scale success requires similar commitment from more people and institutions. Greater exchange within the South between the more and less scientifically proficient countries is a case in point. ICTP has established such links by creating networks, cooperative programs, regional schools, and affiliate centers in the South. Recent efforts by Brazil, China, and India to provide fellowships to promising scientists under a program administered by ICTP's sister organization, the Third World Academy of Sciences (TWAS), suggest that programs for South-South cooperation are finally taking off.

The involvement of scientific institutions in the North is the next crucial element. Here the goal should not be the transfer of technology, but the creation of scientific capacity in each country for generating appropriate solutions for problems involving public health, energy sources, agriculture, ecology, the proper use of environmental resources, and basic education. Other international institutions in Trieste have been working for this goal in diligent partnership with ICTP.

Lasting changes can occur if nations, not just individual scientists, choose to embrace science as an essential part of their national agenda. We must thus move beyond the scientist-to-scientist strategy and become more involved in changing institutions in the developing world. ICTP is increasingly engaging ministries of science and technology in policy discussions, encouraging governments to provide sustainable funding for science. At the same time, we are working in partnership with science institutions in the developing world. This October, ICTP signed an agreement with Brazil's National Council for Scientific and Technological Development (CNPq) to fund four scientific workshops each year in Latin America.

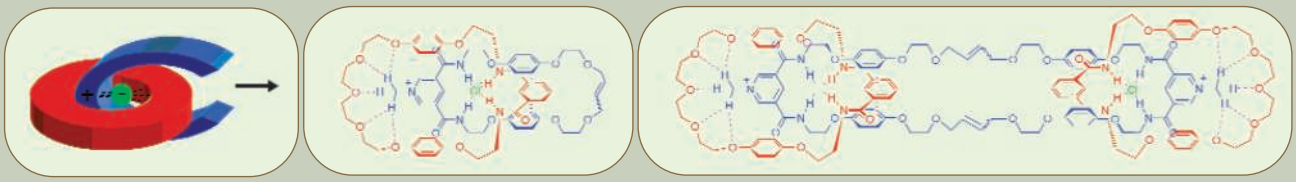
Building scientific capacity is different from instilling a sense of quality. Anchoring quality by providing a well of excellence from which to draw upon will continue to be ICTP's mission and responsibility. That's a full agenda for the next 40 years.

K. R. Sreenivasan

K. R. Sreenivasan is the Abdus Salam Honorary Professor and director of the Abdus Salam International Centre for Theoretical Physics in Trieste, Italy.

EDITORS' CHOICE

edited by Gilbert Chin



Cyclization strategy and products (chloride, green).

CHEMISTRY

Catenane Closure via Chloride

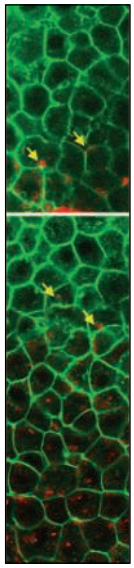
The assembly of interlocking molecular rings, or catenanes, normally relies on some sort of templating mechanism to hold the components together while chemical reactions complete the cyclization. Sambrook *et al*. report on the use of anions as templating agents. They use a catenane precursor and a macrocyclic ring, each of which bears a cleft region that brings two amide groups into close proximity. Binding of a single chloride ion by these four amides holds the precursor onto the macrocyclic ring; this interaction is also stabilized by π-π stacking interactions between hydroquinone groups on both molecules. Ring-closing metathesis cyclizes the precursor, either as a monomer to form two interlocked rings or as a dimer to form a [3]catenane. The [2]catenane product selectively binds chloride anions over acetate and dihydrogen phosphate. — PDS

J. Am. Chem. Soc. 10.1021/ja045080b (2004).

DEVELOPMENT

Restricting Morphogens

During embryonic development, gradients of morphogens and signaling molecules help to define how development proceeds. Scholpp and Brand examined how the gradient of a member of the fibroblast growth factor family, Fgf8, is generated and maintained in the nascent neuroectoderm of living zebrafish embryos. By looking at fluorescently tagged Fgf8 as it spread from its site of origin through target tissue, the authors obtained evidence for a restrictive clearance mechanism in which the factor is cleared from the immediate environment around target cells by endocytosis and subsequent degradation. When endocytosis was blocked, Fgf8 accumulated extracellularly and activated gene expression in more distant target cells, whereas activating endocytosis had the opposite affect, restricting the effective range of Fgf8.



Fgf8 (red) spreads 4 cells away after 1 hour (top) and 12 cells distant after 3 hours (bottom).

Belenkaya *et al*. looked at the movement of another growth factor–related morphogen, Drosophila Decapentaplegic (Dpp), during anteroposterior patterning of the wing. In this system, movement of the growth factor was restricted by binding to extracellular proteoglycans rather than by endocytosis, leading again to a gradient of morphogen response. — SMH

Curr. Biol. **14**, 1834 (2004); *Cell* **119**, 231 (2004).

IMMUNOLOGY

The Cost of Escape

Cytotoxic CD8 T cells (CTLs) begin their assault on the HIV pathogen soon after infection occurs, and the efficiency with which they achieve early control is a deciding factor in the course infection takes. Conversely, the virus defends itself by mutating the epitopes targeted by the CTLs in an attempt to escape recognition. Jones *et al*. explored which characteristics of early CTL responses to HIV corresponded with the subsequent ability to control the viral load.

In an individual showing good viral control, the number and breadth of epitopes recognized by CTLs were relatively large, in contrast to the strong focus of CTLs on a handful of immunodominant epitopes in two individuals exhibiting poor viral control. In these two people, new viruses with numerous CTL epitope mutations appeared soon after infection, suggesting that early selective pressure from CTLs had been countered successfully by the virus. On the other hand, the individual with good viral control carried viruses with far fewer mutations, consistent with the relatively slow emergence of new escape mutants in the months after the acute phase of infection. Early control thus appears to be determined by broad recognition of multiple viral epitopes, increasing both the opportunity for viral detection by CTLs and the potential cost of escape mutations to intrinsic viral fitness. — SJS

J. Exp. Med., **200**, 1243 (2004).

BIOPHYSICS

Unraveling the Knitted Sleeve

The surroundings in which membrane proteins reside consist of a hydrophobic interior (the fatty acid tails of phospholipids), a polar interfacial zone (the phospholipid head groups), and the aqueous compartments on either side of the bilayer. Rather than analyzing the energetics and dynamics of membrane protein insertion in the midst of such heterogeneity, Ganchev *et al*. have resorted to extracting peptides in a model membrane system. A shorter peptide and a longer one, both of which were previously shown to adopt a single-span α-helical conformation in membranes, and two phospholipids, one gel-like and one fluid, were mixed and probed by atomic force microscopy. Pulling (at a range of speeds) resulted in extraction of the peptide, at forces of about 90 pN applied to the gel-like mixture and only 60 pN for the more fluid membrane. A closer look at the resistance to extraction suggests that it arises primarily from the energy required for

CONTINUED ON PAGE 1263

CREDITS: (TOP) SAMBROOK *ET AL., J. AM. CHEM. SOC.* 10.1021/JA045080B (2004); (BOTTOM) SCHLOPP AND BRAND, *CURR. BIOL.* **14**, 1834 (2004)

Published by AAAS

unwinding the first turn of the helix and dragging these residues from the hydrophobic interior into the interfacial region. — GJC

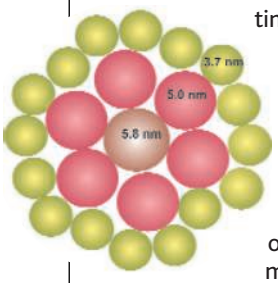
Biochemistry 10.1021/bi048372y (2004).

MATERIALS SCIENCE

A Brighter Future by Working Together

Both metal nanoparticles and semiconducting nanowires have interesting optical and electrical properties, but what happens when they are coupled together? Lee *et al.* try to answer this question for a collection of CdTe nanowires that are complexed with Au nanoparticles using the biotin-streptavidin ligand-receptor pair to connect the two together.

When these components were mixed in solution, the authors observed a fivefold increase in the peak luminescence intensity and a blue shift of the spectra that developed gradually with time. Surprisingly, as the intensity increased, the photoluminescence lifetime decreased, which is in contrast to normally observed trends. The authors interpreted their observations within a model in which the Au nanoparticles form a coaxial shell around the nanowires. They find that the gold particles generate an electro-magnetic field that stimulates photon emission from the nanowires, in a process that is reminiscent of surface-enhanced Raman scattering. This effect is not due to individual nanoparticle-nanowire interactions but instead to the collective effect of the aggregated metallic nanoparticles. — MSL



Cross-section (with diameters in nm) showing the nanowire, streptavidin-biotin linker, and nanoparticle.

CREDITS: LEE *ET AL.*, *NANO LETT.* 10.1021/NL048669H (2004)

Nano Lett. 10.1021/nl048669h (2004).

GEOLOGY

Residence Time

Agricultural and industrial activity has increased the amount of N added to rivers far above natural levels. This N, added mostly as nitrate, is a major pollutant that contributes to eutrophication and produces anoxia in water bodies of all sizes; it also is a source of the greenhouse gas nitrous oxide (N_2O). The magnitude of the impact of riverine N is hard to judge, however, because of large gaps in our knowledge about its removal during transport through the river system.

Donner *et al.* use an aquatic transport model to investigate in-stream N removal and N_2O emissions in the Mississippi River system and how they may be affected by interannual climate variability. Their results show that the fraction of N removed in the river system can vary by nearly a factor of 2, with a threefold range in the associated N_2O emissions, depending on precipitation. The lowest fraction of N removal and the greatest N_2O emissions occur in the wettest years, when river flow is greatest and the residence time of the water in the rivers is shortest. — HJS

Geophys. Res. Lett. **31**, L20509 (2004).

HIGHLIGHTED IN *SCIENCE*'S SIGNAL TRANSDUCTION KNOWLEDGE ENVIRONMENT

Science's stke www.stke.org

Calcium Signals from the Mitochondria

Xu *et al.* used human cell lines that expressed inducible nitric oxide synthase under the control of regulated promoters to investigate the effects of inhibiting mitochondrial respiration with nitric oxide (NO). NO, acting independently of soluble guanylate kinase activity, stimulated expression of glucose-regulated protein 78 (Grp78), an endoplasmic reticulum (ER)–resident chaperone protein whose expression is enhanced as part of the ER stress response. NO produced an increase in the amount of the soluble transcription factor p50 ATF6, which is generated through a calcium-dependent process involving regulated intramembrane proteolysis. NO-dependent stimulation of p50 ATF6 production and of Grp78 expression was attenuated in cells depleted of intracellular calcium, and both an intracellular calcium chelator and cyclosporin A (which interferes with mitochondrial calcium signaling) reduced NO-dependent ATF6 cleavage and prevented the NO-dependent increase in Grp78. Thus, the authors propose that NO-dependent inhibition of mitochondrial respiration affects calcium signaling between the mitochondria and the ER, thereby stimulating production of p50 ATF6 and the expression of genes involved in the ER stress response. — EMA

Nature Cell Biol. **6**, 1129 (2004).

Published by AAAS

NetWatch

edited by Mitch Leslie

IMAGES

Keeping an Eye on Coral

An azure reef rings two of the Society Islands in the South Pacific (below). A new archive can help researchers monitor this and other coral reefs and study how their structures differ in various regions. The library, a collaboration between NASA and the University of South Florida, holds more than 1400 images captured by the Landsat 7 satellite between 1999 and 2003. By clicking on a world map, marine biologists and other researchers can zoom in on a particular reef and download close-up photos. The shots provide baseline data on location and size that are missing for many reefs, making it easier to track changes such as declines that might result from global warming and pollution.

seawifs.gsfc.nasa.gov/cgi/landsat.pl

TOOLS

Worm Genomics Sampler

For their size, parasitic nematodes are disproportionately destructive, ruining more than $80 billion worth of crops around the world each year and causing diseases such as filariasis and trichinosis. Nematode.Net, hosted by Washington University in St. Louis, Missouri, supplies tools for analyzing genomic data from a long list of mainly parasitic worms. The site corrals more than 240,000 expressed sequence tags (ESTs), DNA snippets that can help researchers pinpoint which genes a worm carries. Using NemaBLAST, parasitologists, molecular biologists, and drug designers looking for a worm's weaknesses can hunt for particular sequences in more than 20 nematode species. A search tool lets you view clusters of overlapping ESTs, providing a clear picture of a nematode's genetic endowment.

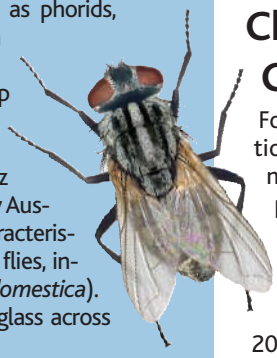
www.nematode.net/index.php

RESOURCES

Flies in the Web

Minute midges, nettlesome gnats, blood-hungry mosquitoes, gangly craneflies. These examples illustrate just a tiny fraction of fly diversity. The Diptera Site,* hosted by the U.S. Department of Agriculture, is a good launching point for fly mavens and neophytes alike looking for information on the more than 120,000 species of these insects. Pages here and on linked sites describe morphology, habits, and other details of 16 fly families, such as phorids, some of which burrow into and feast on ants' heads. Specialists can find possible collaborators in the site's directory or fire up the Nomenclator, a database of valid scientific names.

To brush up on your fly anatomy, buzz over to this linked atlas† being developed by Australian fly experts. Clicking on a list of characteristics highlights them on four representative flies, including the housefly shown here (*Musca domestica*). For close-ups, slide the virtual magnifying glass across the screen.



*www.diptera.org

†www.ento.csiro.au/biology/fly/fly.html

DATABASE

Cold Storage

Before you get sick of trudging through snowdrifts and slipping on slick sidewalks, it might be a good time to visit the National Snow and Ice Data Center in Boulder, Colorado. Experts can dig into more than 400 data sets that record everything from Siberian snow depths starting in the late 1800s to Greenland permafrost temperatures from 1967 to 1982. For instance, satellite images dating back to 1989 let you follow the gradual crumbling of the Larsen Ice Shelf in Antarctica, and thousands of photos of glaciers around the world show how many of these features are disappearing.

The "State of the Cryosphere" section summarizes the latest science on how glacier size, snow cover, sea ice, and other frosty variables may reflect climate change. The site also offers a spectacular gallery, where you can browse historical shots of whopping storms, follow life at a Russian polar station, and view examples of snow and ice formations. The Antarctic landscape above shows the wind-hewn shapes known as sastrugi.

nsidc.org

TOOLS

Chemical Safety Calendar

For authoritative information on the toxicity of common chemicals, many experts rely on the Integrated Risk Information System from the U.S. Environmental Protection Agency (NetWatch, 28 March 2003, p. 1957). The new IRIS Tracker allows users to follow updates to these assessments. The reports proceed through 10 stages, from a literature search to internal and external peer review to appearance on the Web site. For each compound, the schedule lists how far the process has advanced and the expected dates for completing future steps.

cfpub.epa.gov/iristrac/index.cfm

Send site suggestions to netwatch@aaas.org. Archive: www.sciencemag.org/netwatch

CREDITS (TOP TO BOTTOM): NSIDC; NASA/USGS; CSIRO ENTOMOLOGY

Published by AAAS

NEWS

This Week

PAGE 1273
A forerunner of the great apes?

1277
A COX-2–estrogen connection

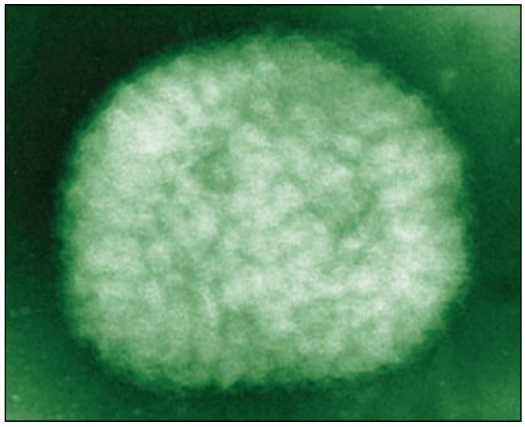
INFECTIOUS DISEASES

WHO Gives a Cautious Green Light to Smallpox Experiments

A World Health Organization (WHO) advisory committee has given its blessing to limited genetic manipulation of the smallpox virus. If the recommendation is accepted by WHO director-general Jong-wook Lee and the World Health Assembly next year, it would mark the first time since smallpox was eradicated that scientists would be allowed to genetically modify the virus.

Once smallpox was wiped out in 1979, the known remaining viral stocks were transferred to two high-security labs at the U.S. Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, and later the VECTOR research center in Koltsovo, Russia. Many involved in the eradication effort pushed for the remaining stocks to be destroyed, but others argued that they should be maintained to allow research on new treatments or safer vaccines in case terrorists or rogue governments have illicit stashes.

A WHO advisory committee must approve any research done with the remaining virus. In a meeting last week, the members gave their initial permission for three new types of work: insertion of a single marker gene into the virus, transfer of one smallpox gene into a related virus, and distribution of very short fragments of smallpox DNA to labs and companies working on diagnostic tests.



Future glow? A WHO committee approved plans to insert green fluorescent protein into the smallpox virus, shown here in a false-color image.

Scientists at the U.S. Army Medical Research Institute of Infectious Diseases want to insert a gene coding for green fluorescent protein (GFP) into the virus to make it easier to screen for new antiviral drugs. A visual assay would make it possible to automate some of the screening tests, making them both faster and safer, says Riccardo Wittek of the University of Lausanne, Switzerland, who heads the WHO committee.

The experiment "has a clear scientific rationale" with little or no chance of accidentally creating a more dangerous virus, agrees molecular biologist Richard Ebright of Rutgers University in Piscataway, New Jersey, who is not a committee member. The review, he says, "is an example of how the process should work."

The panel also decided to relax controls over short stretches of smallpox DNA. According to current rules, anyone wanting to obtain any part of the genome has to apply to WHO for permission, but the committee recommended that fragments of up to 500 base pairs should be freely distributed. Such fragments are too short to code for a whole gene but are used as positive controls in diagnostic tests, Wittek explains.

The committee also decided to permit experiments that would transfer a single smallpox gene—for instance, a gene for DNA polymerase—into a related virus. Scientists have proposed such experiments as a way ▶

ITALY

Academics Protest Plan to End Tenure

NAPLES—Italian academics last week rallied outside Italy's higher education ministry in Rome to show their disapproval of the government's plans to eliminate tenure and increase teaching loads. The rally was the latest in a series of protests against a reform plan the government says would provide much-needed flexibility but which faculty members fear could drive away the country's best young brains.

"This has shown that all the people at all the universities are united," says Piero Tosi, rector of the University of Siena and president of Italy's Conference of Rectors of Italian Universities (CRUI), which views the reforms as an intolerable roadblock to those entering the profession.

In January, Letizia Moratti, Italy's education and research minister, unveiled a draft law that would apply to the majority of the country's 50,000 researchers and professors at its 70 universities. It would replace the current tenured research track with a series of fixed-year contracts at each step along the academic ladder, regular evaluations, and a national qualifying exam. The reforms address widespread claims that the current system is corrupt, with rigged appointments, widespread nepotism, and mismanagement of public resources. These factors, many believe, have fueled a brain drain of the country's best young academic talent.

Although many university professors admit that some of these accusations are well founded, they say the proposed reforms would exacerbate the brain drain by creating intolerable roadblocks to entering the profession. The entire process of winning a tenured slot could take as long as 29 years, scoffs Flaminia Saccà of the University of Cassino, who handles research policies for the Democrats of the Left, the main opposition party in Italy. In addition, says Tosi, the reforms do not address the pressing need to improve evaluation of teaching and research efforts on an individual basis. Academics are also upset by the government's push to double their teaching load, now typically two or three courses a year, and by the government's failure to deliver promised funding increases for research.

The government, which has a solid majority in Parliament, is expected to pass the measures next month, although there could be amendments. Tosi says government officials have agreed to discuss the proposals before the vote, and more protests are planned in order to keep the issue before the public.

–ALEXANDER HELLEMANS

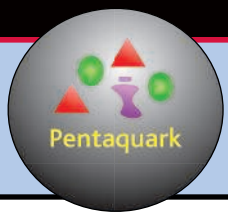
Alexander Hellemans is a writer in Naples, Italy.

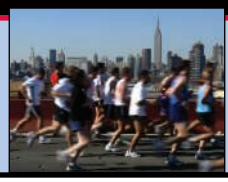
CREDIT: COURTESY OF C. S. GOLDSMITH AND I. DAMON/CDC

Published by AAAS

1278
Probing Iceland's biggest killer

1281
Doubts about an exotic particle





1283
Did running give humans an edge?

to test antiviral drugs without using the smallpox virus itself and would focus on replication genes rather than virulence genes, Wittek says. Even so, such experiments are potentially more troubling than those with GFP, says Jonathan Tucker of the Monterey Institute of International Studies' Center for Nonproliferation Studies in Washington, D.C., because the committee said such work could be done in enhanced biosafety level 3 (BSL-3) laboratories outside of CDC and VECTOR instead of the more secure BSL-4 labs. "My concern is that as the research proliferates, WHO does not have the resources to exercise proper oversight," he says.

But Wittek says that any lab proposing such work would have to go through an extensive review. He said the committee hoped that its approval would speed efforts to find effective treatments for smallpox—one of the goals cited by those who argued for continued research. "It moves you closer to the day when you can destroy the remaining stocks," he says. **–GRETCHEN VOGEL**

FUSION ENERGY

Euro Meeting Holds Key to ITER Project

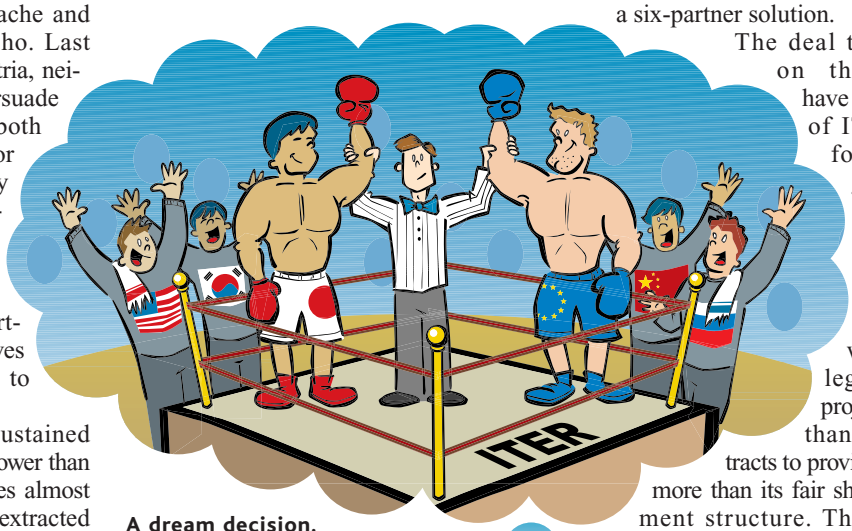
CAMBRIDGE, U.K.—The clock is ticking toward midnight for the fragile coalition trying to build the $6 billion ITER fusion reactor.

This fall research ministers from the European Union (E.U.) set a deadline of 26 November for a decision to begin building the reactor near the French town of Cadarache. But the six partners in ITER are not playing ball: They are currently split down the middle between Cadarache and Japan's proposed site at Rokkasho. Last week, at a meeting in Vienna, Austria, neither the E.U. nor Japan could persuade the other to back down despite both sides claiming to have made major concessions. E.U. officials say they are working to keep the collaboration together, but the ministers' deadline carries with it the implied threat that the E.U. could proceed without the support of all six partners. This week the E.U.'s executives met to prepare recommendations to the ministers.

ITER's goal is to achieve a sustained fusion reaction and generate more power than it consumes. If it works, it promises almost limitless energy, using deuterium extracted from water as fuel and producing little radioactive waste. But first it must be built, at a cost of $13 billion over its expected 30-year life (*Science*, 13 February, p. 940). Last December the United States and Korea decided to back the Japanese site, whereas Russia and China favored Cadarache (*Science*, 2 January, p. 22). Since then each site has been vetted further; delegations have crisscrossed the globe, but neither side has blinked. To break the impasse, the partners have studied the possibility of adding other facilities to the ITER project that would accelerate the move toward commercial fusion power.

In September, frustrated by the impasse, research ministers from the 25 E.U. member states set the 26 November deadline and implied that they would wait no longer on plans to begin construction (*Science*, 1 October, p. 26). The threat of such a unilateral move infuriated the Japanese, who accused the E.U. negotiators of displaying an arrogance that could undermine not just ITER but other international scientific collaborations as well.

In response, Japan quietly began promoting a deal that would minimize the differences between being host and being nonhost. Under the original plan, says Japan's chief negotiator Satoru Ohtake, "being host is like winning the lottery, and being nonhost is like winning nothing." Japan's goal, he explained, was to reach a point at which choosing a site would be "like tossing a coin." But it hasn't fared well, he admits: "I don't think the E.U. ever really imagined being nonhost."



A dream decision. The six ITER partners are looking for a way to anoint both Europe and Japan as winners in the contest to host the reactor.

CREDIT: TIM SMITH

In fact, some E.U. negotiators misinterpreted Japan's overtures as a sign that it was willing to support Cadarache, a position reported erroneously by Reuters news service the day before the 9 November ITER meeting in Vienna. That inaccurate information got the talks off on the wrong foot, says one E.U. source, who added that the meeting ended on friendlier terms after the E.U. delegation restated its support for a six-partner solution.

The deal that the E.U. put on the table would have it contribute 58% of ITER's cost, with four other partners giving 10% each and Japan footing 18%—more than the other nonhosts would give. For its extra money, Japan would get "privileged" status in the project, winning more than 18% of the contracts to provide components and more than its fair share of the management structure. The extra money—contributions would add up to 116% of ITER's nominal cost—would go toward the additional facilities, which Japan would take its pick of. Without showing his hand, Ohtake says that the E.U. proposal "is less generous to the nonhost" than what Japan has offered Europe if the reactor went to Rokkasho.

E.U. officials remain confident that the ITER partners ultimately will embrace the Cadarache site. But continued disagreement remains a possibility, too. If negotiations break down, says one E.U. official, "ITER must still take place." But going ahead with less than six partners "would be a failure," too.

–DANIEL CLERY

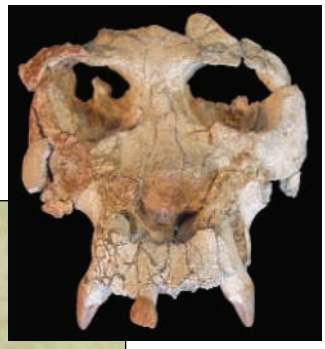
With reporting by Dennis Normile in Tokyo.

Published by AAAS

PALEONTOLOGY

Spanish Fossil Sheds New Light on the Oldest Great Apes

Over the past few decades, paleoanthropologists tracing the human lineage back through time have uncovered a series of increasingly apelike ancestors that date to 4 million to 6 million years ago. Even further back, however, the ancestors of humans and our ape cousins remain mysterious, hidden by a patchy fossil record. Now a Spanish team reports on page 1339 that it has found an exceptionally complete 13-million-year-old fossil that it says is closely related to the earliest members of the great ape family—the large-bodied, long-lived, intelligent clan that includes chimpanzees, orangutans, and humans.



Great-great-grandfather ape. A new fossil (reconstruction, above, and face, inset) may be closely related to the earliest great apes.

CREDITS (TOP TO BOTTOM): S. MOYÀ-SOLÀ *ET AL./SCIENCE*; © AAAS-*SCIENCE*/ILLUSTRATION BY TODD MARSHALL

The new find, from Barcelona, is the most ancient ape to show the upright posture, muzzleless face, and other key traits seen in all living great apes, including people, says paleoanthropologist Salvador Moyà-Solà of the Institut de Paleontologia M. Crusafont in Barcelona, the leader of the team that found the skeleton. In his view, the skeleton is part of the group that gave rise to the living great apes. It illuminates the "total morphological pattern of the early great apes," he says.

Other researchers are delighted with the discovery; paleoanthropologist Carol V. Ward of the University of Missouri, Columbia, calls it an "amazing fossil." But not everyone agrees with the team's interpretations. The ape fossil record of this time, the middle Miocene, is so fragmentary that researchers can reach little consensus on the shape of the ape family tree. "It's a marvelous find, a dream come true," says paleoanthropologist Steven Ward (no relation to Carol) of the Northeastern Ohio Universities College of Medicine in Rootstown. "But the true phylogeny of the great apes is still open to question and will probably not be resolved by this wonderful specimen."

Although dozens of species of Miocene apes have been named, most fossils are fragmentary—a jaw here, an arm bone there. In December 2002, members of Moyà-Solà's team found a canine tooth and an apelike face at a new site near Barcelona, but major excavations had to wait until the summer of 2003. Then the team recovered the ribs, wrist, hands, and vertebrae of a single male individual within an area of about 25 square meters.

The resulting creature, named *Pierolapithecus catalaunicus* after the nearby village of Els Hostalets de Pierola in Catalonia, reveals a mix of apelike and monkeylike traits. Compared to earlier Miocene apes, for example, the face has a much-reduced muzzle resembling the living great apes. The wide and shallow rib cage and details of the vertebrae show that the roughly 30-kilogram creature stood upright, as great apes do. But not on the ground: *Pierolapithecus* was a tree dweller eating fruits and vegetation in a tropical forest. Although it has flexible wrists like those of tree-swinging apes and bipedal humans, it retains the relatively small hands and straight fingers of monkeys, implying that, like them, it sometimes walked on all fours on tree limbs. "From these fossils we have, for the first time in the Middle Miocene, the key diagnostic features of the living apes," together with a large set of primitive monkeylike features, says Moyà-Solà.

The skeletal bones suggest that the early great apes had a somewhat different lifestyle from that of the living ones, says Moyà-Solà. The apelike wrist coupled with a monkeylike hand, for example, suggests to him that our ape ancestors first climbed vertically through the trees and only later began to develop the extensive adaptations for below-branch swinging behavior seen in all living great apes. (Humans lost these adaptations when they came down from the trees.) If the team is right, chimpanzees and orangutans ▶

ScienceScope

EPA Postpones Pesticide Study

The U.S. Environmental Protection Agency (EPA) has suspended a controversial $7 million study of children's exposure to indoor pesticides while it takes another look at the study's design. The Children's Environmental Exposure Research Study (CHEERS) had encountered a flurry of criticism earlier this month, just as it began in Florida (*Science*, 5 November, p. 961).

CHEERS is designed to figure out how children become exposed to indoor pesticides, such as roach sprays. Review boards for the three participating agencies had already blessed the study when the Environmental Working Group (EWG), an advocacy group based in Washington, D.C., criticized EPA for taking $2 million in study funding from the American Chemistry Council (ACC), an industry group. It also questioned whether parents would be adequately warned about the potential dangers of exposing young children to pesticides.

EPA stood by the study but announced on 8 November that it would establish a new panel to review it and report back next spring. "It's great that [EPA] pushed the pause button," says EWG's Richard Wiles. But he still has concerns about industry funding—a topic the new panel isn't expected to address. ACC, meanwhile, says it continues to support the research.

–ERIK STOKSTAD

U.K. Court Orders Animal-Rights Activists to Stand Off

LONDON—The United Kingdom's High Court has ordered animal-rights protesters to stay away from anyone involved in a laboratory construction project at Oxford University. At the university's request, the court last week issued an injunction against seven groups and activist John Curtin, whose protests last July helped shut down work on the $33 million research center (*Science*, 6 August, p. 761).

The injunction was needed to protect researchers, builders, and construction company stockholders from "a small minority of people ... who were undertaking a program of harassment and intimidation" at and away from the construction site, a university spokesperson told *Science*. The order still allows weekly protests at the site by up to 50 people standing outside a 46-meter exclusion zone. It is unclear when work will resume on the project, but university officials say they are aiming to complete the building by late 2005.

–FIONA PROFFITT

Published by AAAS

may have evolved these suspensory adaptations independently, says Moyà-Solà.

Because there are few fossils to put *Pierolapithecus* into context, opinions about its place in the ape family tree vary widely. The team puts it at the key branch point between the great apes and the smaller lesser apes, represented today by the gibbons. Paleoanthropologist David Begun of the University of Toronto, Canada, however, cites facial features that he thinks link *Pierolapithecus* to the African apes, the group that eventually led to chimpanzees and humans, rather than to the Asian orangutans. "I'd put it closer to humans than they would, which makes it even more interesting in some ways," he says.

On the other hand, David Pilbeam of Harvard University thinks that the new skeleton could be even more primitive than the authors suggest. He is not convinced that the characters the team cites—wrist, vertebrae, face, and ribs—indicate an evolutionary link to great apes, and he suggests that the similarities may be due to convergent evolution. "I didn't think the face looked particularly like any living ape. I'm agnostic about the idea that it is part of the group that gave rise to extant apes," he says. "If chimp-orang adaptations are convergent, why believe that *Pierolapithecus* resemblances are not?"

Even as they debate the ramifications of the find, researchers are united in their appreciation of a fossil that is sure to advance the field. "We can't say yet what it all means," says Pilbeam. "But this skeleton is great."

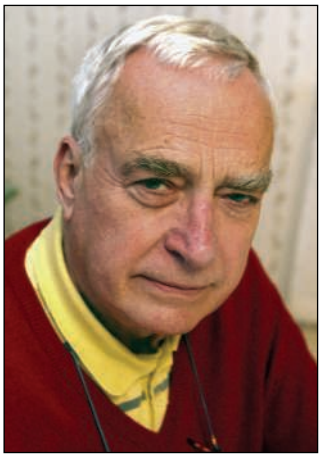
–ELIZABETH CULOTTA

TOBACCO WARS

Research on Secondhand Smoke Questioned

The Philip Morris tobacco company quietly conducted extensive animal research in the 1980s that documented the toxicity of secondhand smoke while arguing publicly that it was safe, according to an analysis, published online last week by *The Lancet*, of thousands of industry and court documents. In a related move, the University of Geneva (UG) has raised doubts about more than 3 decades of tobacco-smoke studies authored by a retired UG environmental-medicine professor who coordinated research for Philip Morris. His failure to disclose that he was a "secret employee of the tobacco industry," according to a UG faculty commission, tainted his research.

The allegations in *The Lancet* drew an immediate response from Philip Morris's parent company, Altria of New York City. It issued a statement saying the charges are "highly distorted and misleading" and that the company has successfully defended against similar allegations in U.S. tobacco litigation. The retired UG professor at the center of the storm, Ragnar Rylander, who is now a professor emeritus at the University of Göteborg, Sweden, maintains that his science was independent and that he is the victim of an antitobacco witch hunt.



Smoking gun? Critics of Ragnar Rylander cite evidence of his work for a tobacco company.

The *Lancet* study grew out of efforts by two antismoking activists, Pascal Diethelm, president of the Swiss antismoking group OxyRomandie, and Jean-Charles Rielle of the Swiss smoking-prevention group CIPRET-Genève. They mined an online database of millions of documents Philip Morris released as part of a 1998 legal settlement with the state of Minnesota. Diethelm, who worked as an information-technology officer for the World Health Organization's Tobacco Free Initiative, knew that Rylander had authored studies exonerating secondhand smoke. A search in the Philip Morris database flagged 16,000 documents in which Rylander's name appeared, including confidential company memos and scientific reports, financial records, and a company consulting contract with Rylander.

In 2001, the two Swiss activists publicly denounced Rylander's work on tobacco smoke as "an unprecedented scientific fraud." Rylander sued Diethelm and Rielle for libel in a Swiss court and won. Diethelm and Rielle appealed; in December 2003 a Swiss appeals court reversed the lower court's decision on grounds that the charges were true. Meanwhile, UG created a faculty fact-finding commission to investigate the charges on its own.

The *Lancet* paper involves many of the players in these battles. Its authors—Diethelm, Rielle, and Martin McKee, a public-health physician at the London School of Hygiene and Tropical Medicine who testified in defense of Diethelm and Rielle at their trial—reported finding more than 800 unpublished studies on secondhand smoke completed between 1981 and 1989 at a Philip Morris facility called the Institut für Industrielle und Biologische Forschung (INBIFO) in Cologne, Germany. In one key 1982 study in rats, INBIFO researchers showed that sidestream smoke, which drifts from lit cigarettes, caused severe damage to the nasal epithelium and abnormal cellular alterations called metaplasia sometimes associated with cancer and was up to four times more toxic than the direct smoke sucked from a cigarette. The data were not published.

The *Lancet* authors maintain that Philip Morris created INBIFO from the start to learn about the effects of tobacco smoke but concealed the work to reduce liability. For example, the authors say, the company situated the lab in Germany instead of the United States, funded it through a Swiss subsidiary, and told few employees about the tobacco-smoke research. The company contracted with Rylander to serve as an intermediary between INBIFO and Thomas Osdene, a Philip Morris executive responsible for research and development.

According to the UG report, released in French on 6 September and in English on 29 October, Rylander also organized industry-controlled symposia that excluded researchers who believed secondhand smoke was harmful and failed to identify himself as a tobacco company consultant in letters to the U.S. Environmental Protection Agency downplaying the toxicity of secondhand smoke.

Last week, UG, acting on advice of the faculty commission, wrote to three journals in which Rylander had published—*The European Journal of Public Health*, the *Archives of Environmental Health*, and the *International Journal of Epidemiology*—warning that "Rylander's work reflects his position as an industry agent rather than a free scientist." From now on, UG researchers will be barred from accepting tobacco-industry grants or contracts, says André Hurst, the university's president.

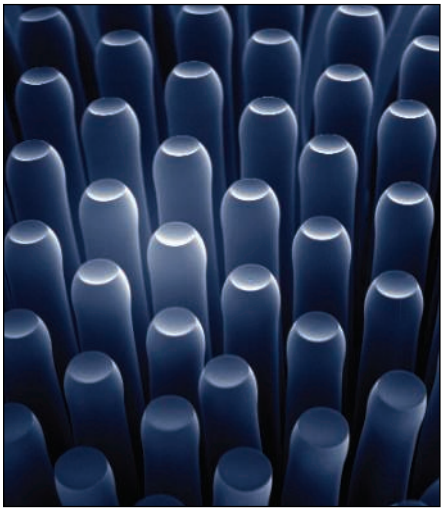
Rylander argues that UG freely accepted Philip Morris's funding for decades and that the authors of the *Lancet* paper quoted from the tobacco documents out of context to "support their message." Furthermore, he says that his science was independent, that he was merely INBIFO's "scientific adviser," and that he had "no insight into their funding." Quoting the Swedish diplomat Hans Blix, Rylander concludes: "If you believe in witches, and you look hard enough, you'll find them."

–DAN FERBER

MATERIALS SCIENCE

Key to Cheaper, Better Nanotubes Comes Out in the Wash

Since their discovery 13 years ago, carbon nanotubes have been nanotechnology's poster child. The tiny straw-shaped molecules are stronger than steel, flexible, and conductive. Researchers have pitched them as the right stuff for everything from chemical sensors and drug-delivery agents to wires for nanoscale computer circuitry and even the building blocks for an elevator extending into space. Their cost, however, is a bit of a problem: At $500 per gram, nanotubes are more than 30 times as expensive as gold. But that price may soon be on its way down.



CREDIT: K. HATA *ET AL.*/*SCIENCE*

Aquaculture. New water-based technique can grow luxuriant columns made of nanotubes.

On page 1362, Japanese researchers report that by simply adding a little water vapor to a standard nanotube production scheme, they've hit upon a new, highly efficient way to grow nanotubes. If the approach can be scaled up, it could significantly drop the price of nanotubes, opening the door to new commercial applications. The team also reports that the technique makes it straightforward to create macroscale sheets, pillars, and other shapes out of nanotubes, which could become the starting materials for novel types of electronic devices. "The results are quite remarkable and will lead to much follow-up," says Hongjie Dai, a chemist and nanotube expert at Stanford University.

In 1991, Japanese physicist Sumio Iijima discovered that nanotubes had grown on the cathode of an arc discharge machine used to make spherical, all-carbon molecules called fullerenes. The machine, which blasts a target of graphitic carbon with a jolt of electricity, turns out a jumble of tubes and soot. Today, most nanotube makers grow their minuscule tubes with the help of tiny nanosized catalyst particles that seed the growth of the tubes inside high-temperature vacuum chambers. The main drawback to this approach is that the resulting tubes wind up contaminated with catalyst particles, which must then be removed through chemical reactions.

In recent years, Iijima, now at the National Institute of Advanced Industrial Science and Technology in Tsukuba, and colleagues have focused on a simple nanotube manufacturing technique called chemical vapor deposition, in which hydrocarbon gases are fed into a superheated chamber containing nanoparticle catalysts. Like other groups, Iijima and his colleagues found that after only about 1 minute of operation, virtually all of the catalysts stopped working. The researchers knew that the high heat broke apart the hydrocarbons, creating a vapor of carbon atoms that link together to form the tubes. The trouble is that the tube must start growing correctly from the catalyst right from the start. Yet in most cases carbon atoms cover the catalyst particles with an amorphous coating that prevents nanotubes from taking shape.

Other researchers had found that they could remove the amorphous carbon simply by adding pure oxygen. But it works a little too well and quickly oxidizes—or burns—the growing nanotubes. "So we figured we need a weak oxidizer that will not damage the carbon nanotubes," says Kenji Hata, the physicist who led the current effort. The group decided to look at water, Hata says, because water readily reacts with carbon to create carbon monoxide and molecular hydrogen. Hata found that when he tuned his apparatus to add about 100 parts per million of water to ethylene and other inert carrier gases, the water reacted with the amorphous carbon from the catalyst particles but didn't damage the growing nanotubes. As a result, virtually all of the catalyst particles remained active and quickly produced a forest of nanotubes growing up from a surface. And because of the high efficiency of the growth process, the resulting crop of nanotubes ends up nearly free of catalyst contaminants.

By starting with catalysts patterned in circles and lines, the researchers grew both pillars and sheets of nanotubes. Because nanotubes have such unique optical, electrical, and thermal properties, patterned tubes may enable researchers to make devices such as optical filters and arrays of electron emitters for flat-panel displays, Hata says.

–ROBERT F. SERVICE

ScienceScope

NIAID Tackles Flu Genomes

Hoping to spur the field of influenza research, the National Institute of Allergy and Infectious Diseases (NIAID) this week announced a new flu genome sequencing project. The $2-million-plus effort will crank through the sequences of thousands of human and avian influenza viruses and deposit them in GenBank, the public DNA database.

Because flu viruses constantly mutate, a new vaccine has to be designed each year for flu season. Having many more sequences on hand will help researchers explore why certain strains are more virulent and improve vaccines and drugs, NIAID officials say. "There's not a lot of sequence out there in the public domain," says the agency's Maria Giovanni. Researchers can also use the data to study how readily a human virus will combine with an avian flu strain, such as the H5N1 strain in Asia, and potentially touch off a global pandemic. The project—part of a broader NIAID microbial sequencing initiative based at The Institute for Genomic Research in Rockville, Maryland—will include collaborators such as flu expert Robert Webster's lab at St. Jude Children's Research Hospital in Memphis, Tennessee.

In another push to prepare for a pandemic, public health experts, government officials, and companies met last week for 2 days at the World Health Organization in Geneva. They called for governments to put up more money for pandemic vaccine development.

–JOCELYN KAISER AND GRETCHEN VOGEL

Infusion for Gulf War Sudies

In a move sure to spur debate, the Department of Veterans Affairs (VA) will spend up to $15 million over the next year on research into Gulf War illnesses, with an emphasis on the role of neurotoxins. The decision, announced on 12 November, follows a key recommendation made by a VA advisory panel that examined ailments arising from the 1990–91 Gulf War (*Science*, 1 October, p. 26).

The panel, chaired by former Defense Department official and Vietnam veteran James Binns, found a "probable link" between the symptoms experienced by Gulf War veterans and toxins that affect the nervous system, such as sarin gas and pesticides. Other committees, in particular those appointed by the Institute of Medicine (IOM), have rejected the neurotoxin hypothesis.

Harold Sox, editor of the *Annals of Internal Medicine* and a member of an IOM Gulf War committee, says the new studies aren't likely to settle the issue because researchers lack good epidemiologic information on what Gulf War troops were exposed to on the battlefield.

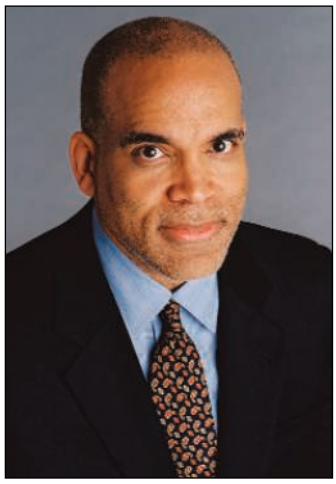
–JENNIFER COUZIN

Published by AAAS

NIH ETHICS

Staff Scientists Protest Plan to Ban Outside Fees

In an unusual collective dissent, more than 170 intramural scientists at the National Institutes of Health are protesting a proposed ban that would prevent them from being paid to advise or speak at institutions that receive funds from NIH. In a letter last week to NIH Director Elias Zerhouni, the scientists, many of them lab and section chiefs, told Zerhouni that a ban on so-called honoraria and other payments is an "error" that risks turning NIH scientists into "second-class citizens in the biomedical community."



Squeaky clean. NIH's Kington aims to avoid the appearance of a conflict.

The letter reflects growing frustration with an ongoing crackdown on consulting, which intramural scientists have mostly endured in silence until now. Many feel that the reforms have gone too far: "There's a tremendous amount of unhappiness," says one lab chief who, like several other signers, declined to be quoted by name.

The ethics overhaul began after the *Los Angeles Times* reported large industry payments to several NIH officials, sparking a congressional investigation and a stringent review of all existing outside activities by NIH. A blue-ribbon panel this spring advised Zerhouni that paid industry work should be banned for top officials and those overseeing grants but permitted for intramural scientists, who are not involved in grant decisions. The panel also deemed it "important—even essential" that NIH continue to allow "reasonable" payments for speaking, writing, and teaching. In the past, NIH routinely approved such lectures, as long as the employee discussed work published at least a year before.

But after Congress pressed for further reforms, Zerhouni announced in June that all paid consulting by intramural scientists for grantee institutions—including speaking—should be banned. NIH Deputy Director Raynard Kington explains that being paid to speak poses the "appearance" of a conflict because intramural scientists "are privy to information about the scientific direction of the agency" that could potentially give the grantee institution an unfair advantage. After talking to legal experts, NIH is also now concerned that if employees discuss their government work at all, they could be using public office for personal gain, Kington says.

The 172 staff scientists who endorsed the 8 November letter disagree. Because intramural scientists are not involved in awarding grants, "there can be no conflict of interest," states the letter, which was initiated by clinical center ethicist Ezekiel Emanuel. Banning activities that are "an essential part of free academic discourse" simply to allay public concerns "seems unjustified," the writers say. The letter also says NIH staff should be able to "very modestly augment" their "low salaries" with these payments, and that barring them "will further erode" NIH's ability to recruit and retain good scientists.

The chief concern, some letter signers say privately, is that because rules have become so restrictive, even being reimbursed for travel to a university for official duty could be disallowed. (Kington says this is not the case.) Others are concerned about NIH's plans for a 1-year moratorium on all industry consulting (*Science*, 1 October, p. 27) and see the honoraria ban as a starting point for a broader discussion. "Many of us feel this is the first chink in the armor to attack," says virologist Malcolm Martin.

Some contend that NIH recruitment efforts could be suffering already. Neuroscientist Bruce McEwan of Rockefeller University in New York City says the NIH intramural program is attractive at a time when extramural grants are getting tighter, but on the other hand, the agency's salary cap and limits on outside income "make the situation very difficult" for academic scientists.

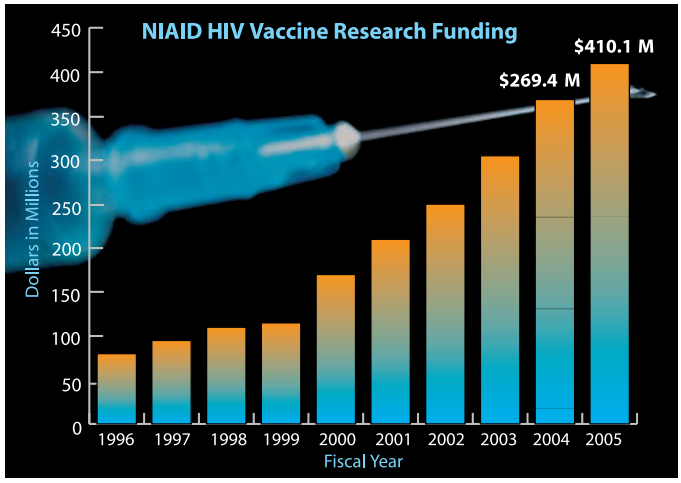
Kington says there's no evidence that scientists are being driven away from NIH but that the 1-year pause in industry consulting will allow NIH to "measure impact." Meanwhile, Zerhouni and Kington plan to meet with the scientists who signed the letter on 29 November.

–JOCELYN KAISER

AIDS VACCINES

The First Shot in a Highly Targeted Strategy

For biologists, it's a new way of doing business. Last week, the National Institute of Allergy and Infectious Diseases (NIAID) unveiled a highly targeted strategy to tackle immunological puzzles bedeviling AIDS vaccine research. Rather than rely on individual researchers to submit grant proposals, NIAID is planning to bankroll a collaboration called the Center for HIV/AIDS Vaccine Immunology (CHAVI) to focus on problems already identified as key obstacles to the development of an effective AIDS vaccine. The proposed venture could pump at least $300 million into the field over the next 7 years. And more such initiatives are on the way.



Booster shot. NIAID says the new center will not take funds from its existing HIV/AIDS vaccine program, which has been growing steadily.

CHAVI is the first tangible outgrowth of the Global HIV Vaccine Enterprise, an unusual international effort to set priorities for AIDS vaccine research and development and to organize the field to tackle the critical problems. First described in *Science* last year (27 June 2003, p. 2036) and endorsed by leaders of the wealthy "G8" nations at their summit in June, the fledgling enterprise has held several meetings of leading AIDS researchers to draft a "scientific blueprint" laying out goals and strategies. The idea for ▶

CREDITS (TOP TO BOTTOM): NIH; DHHS/NIH/NIAID/DAIDS

Published by AAAS

CHAVI comes from these discussions, says Peggy Johnston, who heads AIDS vaccine development at NIAID. The institute decided to "to get ahead of the game" by taking action even before the blueprint is published, she says. "This is one of [NIAID's] highest priorities."

NIAID outlined the plans for CHAVI at a workshop last week. The institute intends to award a 7-year contract to a group of investigators—possibly from several institutions, including ones outside the United States. The collaboration will focus on unraveling the immune responses that have protected monkeys from the AIDS virus in a few well-known experiments that no one has ever satisfactorily explained (*Science*, 28 June 2002, p. 2325). It will also investigate why some people who are repeatedly exposed to HIV remain uninfected, how the virus changes soon after it establishes an infection, and what the comparative strengths and weaknesses are of various vaccine approaches. CHAVI will need a full-time director, someone willing to give up grants that do not directly relate to the center's research. "It requires a mentality that has not been pervasive in the AIDS community," says NIAID Director Anthony Fauci.

Researchers must submit their proposals for CHAVI by 23 February, and NIAID plans to make the award before the 2005 fiscal year ends on 30 September. Fauci does not yet know where the money will come from to fund what could become one of NIAID's most expensive projects but promises that CHAVI will not siphon funds from the institute's existing AIDS vaccine budget (see graph).

The general idea for CHAVI is drawing praise from the AIDS research community. "All of us here love the proposal," says J. Michael McCune of the Gladstone Institute of Virology and Immunology in San Francisco, California, one of about 50 researchers who attended the meeting (which was broadcast live on the Internet). "It's a very important, great first move," adds Lawrence Corey of the University of Washington, Seattle.

Other significant initiatives could come in the next few months from the Global HIV Vaccine Enterprise. Jose Esparza, a senior adviser to the Bill and Melinda Gates Foundation, who serves as the enterprise's interim secretariat, says a final draft of the scientific blueprint is on his desk, and he expects it to be published "very, very soon." The Gates Foundation also plans in the next few months to announce its own grants that tie into the blueprint. "Our intention is to provide a very significant level of new resources," says Richard Klausner, head of the foundation's Global Health Program. **–JON COHEN**

MEDICINE

Estrogen's Ties to COX-2 May Explain Heart Disease Gender Gap

Estrogen and the enzyme COX-2 have more in common than their ability to spark enthusiasm tainted by controversy. It turns out that one recruits the other in the blood vessels of mice. This discovery raises the possibility that COX-2 inhibitors—including Vioxx, which was yanked off the market in September because of cardiovascular risks—might be particularly hazardous to females, especially younger ones whose bodies are still churning out estrogen. The work may also help explain how estrogen production in premenopausal females benefits the heart, a hot-button issue ever since hormone replacement therapy in older women was found to boost heart disease.

Unprotected. Plaques (red) spread in the arteries of a female mouse lacking a key receptor.

CREDIT: K. M. EGAN *ET AL.*/*SCIENCE*

"It opens a new option for re-evaluating phenomena which we have been wondering about for decades," says Kay Brune, a pharmacologist at the University of Erlangen in Germany. Although he and others caution about extrapolating the findings to humans, Brune nonetheless calls them "potentially of enormous clinical significance."

The study, published online by *Science* this week (www.sciencemag.org/cgi/content/abstract/1103333), was led by Garret FitzGerald, a pharmacologist and cardiologist at the University of Pennsylvania in Philadelphia—and an outspoken critic of COX-2 inhibitors. He and his colleagues were curious about a fatty acid called prostacyclin, which is produced by COX-2 and whose synthesis inhibitors such as Vioxx block. Cardiologists consider the loss of prostacyclin a likely culprit behind Vioxx's woes (*Science*, 15 October, p. 384).

FitzGerald's group created mice genetically susceptible to atherosclerosis and lacking the prostacyclin receptor. The scientists were intrigued to see that in these animals there was no gender gap in heart disease—a divergence long observed in both people and mice in which younger males are at higher risk than younger females. "Females caught up to males," developing atherosclerosis just as quickly, says FitzGerald. A closer look revealed that without the prostacyclin receptor, female mice were highly susceptible to oxidative damage from free radicals, which boost plaque formation in arteries.

Then the team sought to bring estrogen into the picture, turning to mice whose ovaries had been removed and who were given supplemental estrogen. The hormone supplements, FitzGerald and his colleagues found, increased prostacyclin biosynthesis and depressed oxidative stress; both effects were tracked in urine samples.

This suggests that in premenopausal females, estrogen, acting through one of its receptors, stimulates COX-2 production. That boosts prostacyclin, which in turn protects the heart from atherosclerosis, says FitzGerald. This may explain much of the gender gap in heart disease, because that estrogen-driven pathway would be much weaker in males.

But as wary researchers know, mice aren't people. For decades, based on studies in mice and other animals, millions of menopausal women took hormone supplements believing the drugs would stave off heart disease. Then, in 2002, the Women's Health Initiative (WHI) reported that women on hormones were more likely to suffer heart problems than those on placebos.

One question raised by the new study is whether women in WHI taking COX-2 inhibitors or general nonsteroidal anti-inflammatory drugs influenced the study's outcome. Women in the study on aspirin didn't appear to be at higher risk, says Richard Karas of the molecular cardiology research institute at Tufts University in Boston. But FitzGerald is beginning to comb through the hormone data from WHI and other studies to take a second look. His team's findings, he notes, could be much more applicable to younger women; although the published COX-2 inhibitor studies haven't noted gender gaps in heart risks, most participants were older.

Two other COX-2 inhibitors, Celebrex and Bextra, remain on the market, and scientists are trying to determine whether they share Vioxx's risks. FitzGerald's work "raises the question of whether there are gender-specific risks related to the COX-2 inhibitors," says Karas. If it holds up, he adds, it also suggests that the cardiac hazards "are not some weirdo side effect of Vioxx that's not true in the other" drugs.

–JENNIFER COUZIN

Published by AAAS

The more researchers learn about the unheralded Laki eruption of 1783, the more they see a need to prepare for a reprise that could include fluoride poisoning and widespread air pollution

Iceland's Doomsday Scenario?

SKAFTÁRTUNGA, ICELAND—Hildur Gestsdóttir shovels a heap of fine black soil onto a growing mound beside the unmarked grave, grateful for a breeze from a nearby glacier that's taking the edge off the strong summer sun. "It's a lovely day for gravedigging," a member of her team remarks. Hildur agrees: "Conditions are perfect."

Hildur ought to know, having exhumed about 50 skeletons to date with the Institute of Archaeology in Reykjavik. Usually she's after the remains of Vikings, who settled the island 1000 years ago, or later medieval inhabitants. This grave is much more recent, dating from the late 18th century. Although the period is not her forte, the skeleton beneath Hildur's feet on Búland farm could well be a researcher's treasure, offering clues to why the eruption of the nearby Laki fissure in 1783 was so deadly. One of the largest and least appreciated eruptions in recorded history, Laki killed 10,000 Icelanders—roughly one in five—and recent studies suggest that its billowing plumes led to extreme weather and extensive illness that may have claimed thousands more lives in Britain and on the European continent.

"It's hard to fathom the impact of Laki," says volcanologist Thorvaldur Thordarson, a leading expert on the eruption. A similar blast in modern times would pump so much ash and fumes into the upper atmosphere that the ensuing sulfuric haze could shut down aviation in much of the Northern Hemisphere for months, Thordarson and Stephen Self of Open University in Milton Keynes, U.K., argued last year in the *Journal of Geophysical Research*.

"It's not a matter of if but when the next Laki-like eruption will happen" in Iceland, says Thordarson, who splits his time between the University of Iceland and the University of Hawaii, Manoa. "We certainly don't want to be here when another Laki-type event hits," adds Self. Offering a tame glimpse of what the future may hold, the brief eruption of Iceland's Grímsvötn volcano earlier this month led to the cancellation or rerouting of transatlantic flights. Still, volcanologists say, the odds of a full-blown fissure eruption in this century are low.

By examining presumed victims of Laki, Hildur and her colleagues, including project leader Peter Baxter, a medical researcher at the University of Cambridge, U.K., are testing a thesis that fluoride in Laki's emissions poisoned people directly and may account in part for the high death toll. "It was the greatest calamity to affect Iceland since human occupation began there," says Baxter.

Putting her back into it. Archaeologist Hildur Gestsdóttir digs up graves on Búland farm in search of Laki's victims.

During the eruption, an estimated 1 million tons of hydrofluoric acid were deposited over Iceland, contaminating the country's food and drinking water supplies. Icelanders who lived through the eruption noted that sheep and other livestock developed knobbly protrusions from their bones that were clearly visible under the skin—a telltale sign of fluorosis. Baxter's team is the first to exhume presumed victims of Laki to look for abnormal bone growth and high levels of fluoride that could well have led to fatal poisoning in people during the later months of the eruption.

If they are right, Iceland's fissure eruptions may be much more dangerous than scientists had supposed. And this realization implies that civil-defense planners need strategies for the next Laki-like event. "It's important to consider what the next one is going to do, and how we can prepare for it," says Clive Oppenheimer, a volcanologist at the University of Cambridge.

Fire and brimstone

Laki, a fissure in the basalt lava fields of Iceland's southeastern fringe about 50 kilometers north of Búland, embraces 140 volcanic vents that seem to march in a neat row, 27 kilometers long, toward massive Grímsvötn in the northeast. Hunkered beneath a glacier, Grímsvötn is a restless giant, awakening every 10 to 15 years on average. Its eruptions, including one that began on 1 November and lasted 5 days, unleash torrents of glacial meltwater—awe-inspiring floods called *jökulhlaup*—onto the coastal plains.

Iceland is the only spot on Earth above sea level where fissures, formed by spreading at midocean ridges, are likely to erupt on a titanic scale. Laki-like events happen every 500 to 1000 years or so, although "you can have a group of eruptions in a very short period," says Thordarson. And if fluorine were gold, Iceland would be a fabulously wealthy nation. "Not all magmas are fluorine-rich. Persistent offenders seem to include volcanoes in Iceland and Melanesia," says Oppenheimer.

The last time Laki roared to life, all hell broke loose. Reverend Jón Steingrímsson was an eyewitness who recorded his observations of the 1783 eruption in his *Eldrit*, recently translated into English as *Fires of the Earth*:

Around midmorn on Whitsun, June 8th of 1783, in clear and calm weather, a black haze of sand appeared to the north of the mountains nearest the farms of the Síða area. … That night strong earthquakes and tremors occurred.

Steingrímsson's chronicles of the months-long spectacle are "phenomenal,"

CREDITS: R. STONE/*SCIENCE*

Published by AAAS

says Thordarson. He was the first to describe "Pele's hair"—ash "shaped like threads," Steingrímsson wrote, "blue-black and shiny, as long and thick around as a seal's hair." He also first recorded spatter bombs: blobs of lava hurled into the air that splat "like cow dung" after landing, says Thordarson. Whereas volcanoes like Mount St. Helens and Pinatubo erupt explosively, Laki and its brethren erupt effusively, similarly to the relatively tame lava fountains spilling out of Hawaii's Mount Kilauea. As one of the largest documented effusive eruptions, Laki lasted 8 months, disgorging an estimated 14.7 cubic kilometers of lava, approximately 150 times the average amount for a basalt eruption and enough to cover 580 square kilometers of the island.

Laki cast a deathly pall over Iceland. "The entire country was basically engulfed in volcanic fumes," says Thordarson. "They couldn't even go out fishing; they would get lost in the haze." The prodigious emissions, he says, included an estimated 122 million tons of sulfur dioxide, 7 million tons of hydrochloric acid, and 15 million tons of deadly hydrofluoric acid.

The foul smell of the air, bitter as seaweed and reeking of rot for days on end, was such that many people, especially those with chest ailments, could no more than half-fill their lungs of this air, particularly if the sun was no longer in the sky; indeed, it was most astonishing that anyone should live another week.

Vapors of death

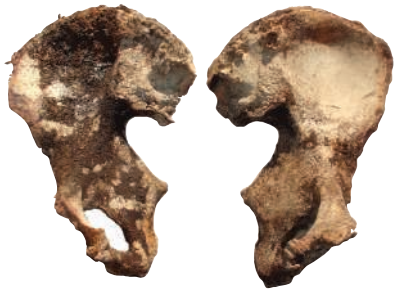
Laki's acrid tendrils reached far beyond Iceland's shores. Scores of published accounts of weather conditions on continental Europe during the summer of 1783 refer to a persistent haze, or "dry fog," that lasted months. Ever ahead of his time, Benjamin Franklin reasoned that the dull sun and blood-red sunrises and sunsets were the result of a volcanic eruption in Iceland.

Piecing together a more detailed picture of Laki from historical documents and from their own computer modeling, Thordarson and Self concluded in the *Journal of Geophysical Research* last year that Laki's eruption columns rose as high as 13 kilometers into the air. "No one had envisioned such explosive powers in a fissure eruption," says Thordarson. The resulting aerosol veil hung over the Northern Hemisphere for more than 5 months.

In a cruel twist, the volcanic haze rolled in just as Europe was wilting in an unusually hot summer. The fumes, some researchers argue, sent thousands of people to an early grave. In a paper in press at the journal *Comptes Rendus*, a team led by John Grattan of the University of Wales in Aberystwyth reports that, according to burial records, there were 25% more deaths than usual between August 1783 and May 1784 in 53 parishes across France. Extrapolating these numbers countrywide, they write, Laki's death toll in France "may be far in excess" of the 16,000 people whose deaths have been linked to air pollution and oppressive heat during the summer of 2003.

Laki writ small. This month's eruption of Grímsvötn led to flight cancellations and sent farmers in Iceland scrambling to shelter livestock from fluorine-rich ash.

Extending Grattan's work in England, Oppenheimer and Claire Witham of the University of Cambridge reported last May in the *Bulletin of Volcanology* that about 20,000 people in England alone succumbed to climate anomalies in the summer of 1783 and the following winter. Scouring burial records of 404 English parishes over a 50-year period spanning the years 1759 to 1808, they found that August–September 1783 and January–February 1784 were especially lethal months. Weather records confirm that the summer of 1783 was notably hot in England.



Volcano victim? Nodules at the top of this pelvis, which once belonged to a blonde-haired woman in her 30s, suggest fluorine poisoning.

The following winter was one of the most severe ever recorded in European annals. Anecdotal reports point to a shortage of firewood throughout Europe, and Europeans were dying in droves during that winter, according to findings published by Grattan and colleagues in the late 1990s. The mean surface cooling in Europe during 2 years following the eruptions was about 1.3°C, according to Thordarson and Self. They blame Laki's aerosols for having disrupted the Arctic "thermal balance."

Oppenheimer acknowledges that it's "a challenge" to make a direct link between Laki and the sharp spike in mortality in 1783–84. "People may have been hit by a cocktail of things," he says. But if Laki were the primary cause, it would be the third most deadly eruption in history, after Tambora in 1815 and Krakatau in 1883.

Divining the bones

A tractor, engine growling, zigzags along a slope above Búland farm, tilting precariously as it turns hay. From the top of the hill you can see for miles: waterfalls fed by glacial runoff plunge through crags in verdant hills, sheep graze near a braided river, and a vast plain of moss-encrusted Laki lava stretches to the horizon.

Near the farmhouse, a cluster of dwarf birch trees mark the boundary of what once was a cemetery. In the 18th century, the graves abutted a church that has long since vanished.

Baxter kneels next to the mound of soil dug from the grave. Death is his forte. Nicknamed by colleagues "Dr. Doom," Baxter has pioneered the study of how volcanoes kill (*Science*, 28 March 2003, p. 2022). "This is remarkably fine ash," he says, sifting it through his fingers. Long ago, fluoride compounds were identified as the culprit responsible for much of the loss of livestock during and after the eruption. Baxter notes that fluoride salts ad-

CREDITS (TOP TO BOTTOM): R. STONE/*SCIENCE*; ALEXANDER H. JAROSCH/INSTITUTE OF EARTH SCIENCES, UNIVERSITY OF ICELAND

Serene desolation. The abandoned graveyard of Ásar church overlooks a vast plain of weathered Laki lava; a skull emerges from excavations at Búland farm.

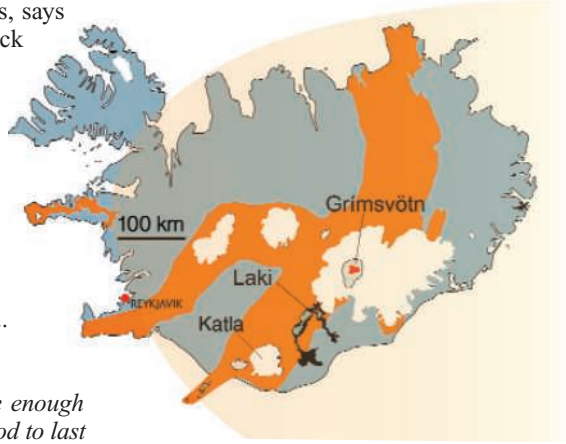
here to ash particles, which in turn would have clung to the vegetation and would have been consumed in prodigious amounts by grazing animals. Even during this month's reawakening of Grímsvötn—part of the same volcanic system as Laki—farmers in areas where the ash fell brought livestock indoors to prevent the animals from ingesting fluoride-laced ash.

During the Laki eruption, "fluorine poisoning was observed all over Iceland" in the form of bone malformations, says Thordarson. "We know the livestock were being poisoned and that within months people started dying," says Hildur. "But no one wondered whether people were also dying from direct poisoning" from contaminated food or water. Indeed, says Baxter, Steingrímsson's accounts of abnormal bone growths in people were long overlooked. The reverend wrote:

Those people who did not have enough older and undiseased supplies of food to last them through these times of pestilence also suffered great pain. Ridges, growths and bristle appeared on their rib joints, ribs, the backs of their hands, their feet, legs and joints. Their bodies became bloated, the insides of their mouths and their gums swelled and cracked, causing excruciating pain and toothaches.

In their pilot study, Baxter and his colleagues have looked for graves in cemeteries at Búland and the nearby Ásar church that were abandoned at the end of the 19th century. Hildur and her colleagues dated the graves according to layers of volcanic ash in the soil. The grave at Búland, for example, was dug shortly after the ash fell from Laki but well before the ash from an 1845 eruption of the Hekla volcano.

Hildur, in a hole that's now more than a meter deep, uses a spade to clear dirt from the coffin lid. She dons surgical gloves and begins removing pieces of the decaying wooden lid. After a half-hour of painstaking work, she exposes the skull to the light of day. Its matted blonde locks are stunningly preserved.



Superhot spot. Laki's 140 vents run diagonally from Katla to Grímsvötn at the center of Iceland's volcanic zones.

Hildur passes the fragile remains to her officemate, Guðrún Alda Gísladóttir, who stows them in Ziploc bags. The pelvis, from a woman apparently in her 30s, has nodules protruding near the top edge. "This is very unusual," says Baxter. "It may well be the result of fluorine poisoning." The pelvis of another presumed Laki victim exhumed in the spring was similarly misshapen. "Two graves, randomly chosen, showing the same changes," he remarks. To trigger such bone growth, "you would have to have really slugged them with fluoride."

The heftiest doses would have come through drinking water, possibly up to 30 or 40 parts per million—as much as 30 times the permissible level today, says Baxter. "It was high enough that you would have felt sick if you drank the water," he says. "But they were in such a terrible state, they had no choice." The Icelanders were already suffering from deficiencies in vitamins C and D. "Then add fluorine," he says. "Nutrient deficiency could have made the population much more susceptible to fluorine poisoning."

In September, bone samples were shipped to the University of Cambridge for testing. There, a team led by Baxter and Juliet Compston is measuring the levels of fluoride and other trace elements, such as arsenic, in ashed bone samples. "It could be a soup of chemicals from the volcano," says Baxter. Georges Boivin of the University of Lyon, France, is now using x-ray diffraction to determine precisely how the fluoride ions were substituted for other minerals in the bone's apatite crystal matrix. Results are due by the end of the year. Baxter hopes that the preliminary findings will lead to a "robust" study involving many exhumations, which could nail the fluoride link.

The next apocalypse

The Laki eruption has been a tragedy lost in time. "People ignored it for so long," says Thordarson.

That's changing. Volcanologists now view Laki as a potent warning, and some are considering what could be done to prepare for a reprise, beyond protecting food supplies and handing out respiratory masks.

Some potential consequences could not have been dreamed of the last time Laki erupted. The atmosphere, laden with

CREDITS (TOP TO BOTTOM): PHOTOS: MUTSUMI STONE/*SCIENCE*; T. THORDARSON/UNIVERSITY OF ICELAND

Published by AAAS

charged particles, would bristle with electricity, possibly interfering with satellite communication. And the plume could well wreak havoc on civil aviation. "How would British Airways deal with its jets being grounded for 5 months?" asks Thordarson. "Planners would be smart to think ahead about how they might deal with such a contingency," says Christopher Newhall, a volcanologist with the U.S. Geological Survey in Seattle, Washington. However, he notes, "the chances of the next one happening in our lifetimes is relatively low."

At the moment, Iceland's fissures do not seem to be up to trouble. "We have not seen potential precursors for an eruption," says Freysteinn Sigmundsson of the Nordic Volcanological Centre in Reykjavik, who serves on the science committee of Iceland's civil defense department. Precursors could include earthquakes, deformation of the earth's crust, or an uptick in geothermal heat.

But volcanic fissures are hard beasts to track. A full-blown fissure eruption would follow an upsurge in magma from reservoirs near the crust-mantle boundary about 10 to 20 kilometers below the surface. Before the next Laki-type eruption, huge volumes of magma need to accumulate—as much as 15 cubic kilometers, roughly the amount generated under all of Iceland over a span of 100 years, Sigmundsson says. Although a strategy for monitoring precursors of such events remains elusive, he says, satellite radar imagery can detect crustal deformation—and thus magma accumulation—as deep as the crust-mantle boundary. "Judging from Laki, we would have 3 to 4 weeks of precursor activity," mainly in the form of earthquakes," Thordarson says.

Yet there are uncertainties galore. High magma pressure at Grímsvötn and Katla—a volcano just to the southwest of Laki—could trigger a failure of the plate boundary between the volcanoes, which in turn could spark a fissure eruption, Sigmundsson says. Civil defense officials will remain vigilant for signs of such an event, he says: "We're following the situation closely."

A Laki-esque eruption could also occur in other volcanic systems in Iceland. Katla's current bout of insomnia is particularly disconcerting. The biggest fissure eruption in recorded in history was that of the Eldgjá fissure just east of Búland and connected via its plumbing to Katla. Over 6 years beginning in 934 C.E., Eldgjá spewed about twice the amount of sulfurous materials into the air as Laki later produced. "Eldgjá had a huge environmental impact and probably stopped settlement of Iceland for some years," says Thordarson. "In that eruption the fissure and Katla volcano erupted simultaneously."

Current scientific interest in Laki and its ilk stems in some measure from a new appreciation for the observations of Steingrímsson, who saw the eruption as a religious apologue that would die with him unless he committed it to paper. As he wrote in his forward to *Eldrit*, "I thought it would be unfortunate if these memories should be lost and forgotten upon my departure." The deformed, fluoride-laden bones that Hildur and Baxter have unearthed may provide another powerful testament to the peril of taking Iceland's fissures lightly.

Thordarson, for one, is intent on persuading colleagues and the general public that Laki is a sleeping giant that cannot be ignored. "We're much better off if we prepare ourselves for the worst-case scenario," he says. "I'm not trying to be a doomsayer. But it could happen tomorrow." –RICHARD STONE

SOURCE: K. HICKS/OHIO UNIVERSITY

High-Energy Physics

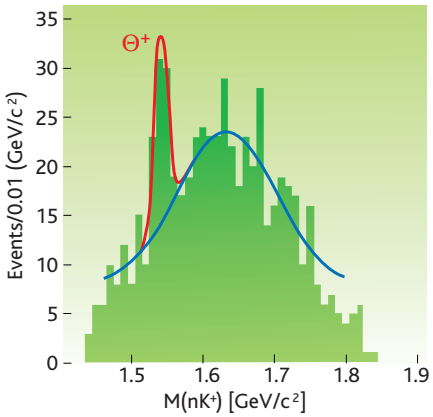
Rara Avis or Statistical Mirage? Pentaquark Remains at Large

Two years after its surprise appearance in debris from a nuclear collision, some researchers suspect an exotic particle may be a will-o'-the-wisp

Two years ago, scientists in Japan and then the United States made headlines around the world by announcing that they had found an unusual particle. This creature, dubbed the Θ^+ (theta-plus), was apparently made of five quarks rather than the two or three quarks that make up all other known quarky matter in the universe. That unique property would make the so-called pentaquark a totally new way to probe the forces that hold atoms together (*Science*, 11 July 2003, p. 153). "It's a fantastic beast—if it exists," says Ted Barnes, a physicist at Oak Ridge National Laboratory in Tennessee.

But the beast might be mythical. Even though a dozen experiments have independently claimed to detect the Θ^+ particle in their data sets, and physicists at the Thomas Jefferson National Accelerator Facility (JLab) in Newport News, Virginia, are trying to corner the particle, some particle physicists are murmuring their disbelief. As negative results and inconsistencies pile up, many scientists suspect that pentaquark aficionados are chasing a phantom.

Edward Hartouni, a physicist at Lawrence Livermore National Laboratory in California, is part of a team that pored through the debris of a billion energetic particle collisions, searching for evidence of pentaquarks. If the particles exist, they should show up on data plots as a huge spike. They are missing. "There is no large peak here," Hartouni says.



Shaky pillar? Data spikes hinting at new type of matter have drawn fire from skeptics.

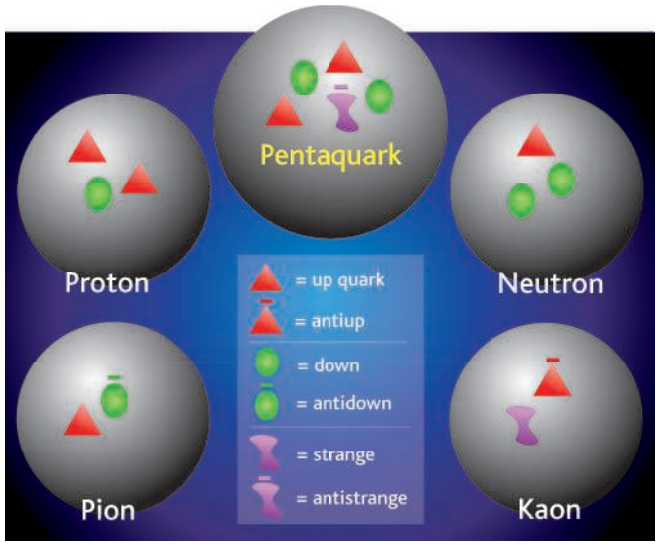
The case in favor

It's hard to believe that something that has been spotted at so many laboratories might be an artifact. Indeed, these laboratories seem to have spotted the Θ^+ in different ways. Some, including the SPring-8 experiment in Japan and those at JLab, zap nuclei with light. Others, including experiments in Russia and Germany, smash mesons or protons or electrons into nuclei. The result in each case seems to be the same: a "peak" in the data that signals the brief life of a five-quark particle—whose mass is a bit more than one-and-a-half times that of the proton—that quickly decays into a handful of smaller particles.

Particle physicists have been finding such peaks for decades. In the spray of debris after a collision, there's a wealth of information as to what happened. Scientists with sufficiently good detectors can look at the tracks of the debris and identify what those particles were. By tracing the tracks backward and seeing how they combine or split apart or kink or curve, physicists can infer what sorts of particles were created in the smashup, how heavy they are, how much charge they carry, and what they decay into. Often, scientists will graph data in a way that shows the number of times that a certain collision yields an event with a giv-

en energy. A lump in that data will often indicate the repeated creation of a particle with a given mass-energy; for example, scientists running the right type of experiment will see a clear peak at 1520 MeV that indicates the creation of a three-quark particle known as the Λ(1520).

The better the detectors are and the more events are analyzed, the more starkly a particle's lump will stand out amid the bumps and wiggles of background noise and statistical fluctuations. According to Kenneth Hicks, a physicist at Ohio University, Athens, and member of the JLab team, lots of experiments have detected a peak that would signal the creation of the Θ^+ particle—a very narrow peak at about 1540 MeV—with good statistical significance. "Some appear to be very significant," he says. "About one dozen experiments have seen it with statistics better than 1 in 1000." Or, more precisely, there have been quite a number of "three-sigma" detections, which are the informal gold standard of statistical significance under many circumstances in particle physics. "Many [detections] are higher: four, five, six sigma," adds Hicks. Five- and six-sigma results are usually considered quite high quality, and one that turns out to be false can wind up being a high-profile embarrassment (*Science*, 29 September 2000, p. 2260).



Oddball. All other known quark-based particles contain either two or three quarks; controversial pentaquark would boast five.

The case against

Even so, some physicists remain unconvinced. For a start, the apparent sightings pose theoretical problems. The narrower the peak, in general, the more stable the corresponding particle, and pentaquarks shouldn't be all that stable. According to theories about the forces that bind nuclei into stable packages, it's difficult to see why a five-quark ensemble, for example, wouldn't rapidly decay into a baryon (with three quarks) and a meson (with two). "It should spontaneously fall apart," says Barnes. "There's no reason for it to stay together." If some unknown mechanism keeps it from splitting into fragments, it must bind the five quarks quite tightly to create the narrow peak.

Although this is difficult to understand, it might be the sign of exotic new physics for theorists to figure out. However, although the peak is narrow—25 MeV wide or smaller, with its apex pinned within 5 or 10 MeV—its center seems to move around. The experiments that claim to have seen the Θ^+ peg its mass-energy at anywhere between about 1525 MeV and 1555 MeV. "That's an extraordinarily large range," says Michael Longo, a physicist at the University of Michigan, Ann Arbor.

"It worries some people, but it doesn't worry me," counters Hicks, who argues that the inherent errors in pinning down masses can account for the differences between the experiments. "It's consistent within the error bars."

More disturbing to skeptics, though, is that a number of efforts to mine old data for signs of the Θ^+ have come up empty. For example, Longo's team reanalyzed the debris of proton-proton collisions at Fermi National Accelerator Laboratory (Fermilab) in Batavia, Illinois, and failed to find any five-quark particles. "One of the things we thought we would be able to see is the pentaquark state Θ^+," he says. "We thought we'd just verify the existence of the state. It just wasn't there." Hartouni's group also looked at collisions at Fermilab, and although they found a nice sharp peak that signaled the existence of the nearby Λ(1540) particle, there was no hint of a spike for the Θ^+ at 1540 MeV. "At 1540, we see essentially no [Θ^+] production," he says.

Hicks calls the Hartouni result "one of the most significant results of a nonobservation." But with so little known about how pentaquarks are produced, he says, it's possible that high-energy collisions like Fermilab's might not create pentaquarks in the same manner as the lower-energy collisions at JLab. Others are less sure. "If, in fact, the JLab results are confirmed, we have a true puzzle," says Hartouni. "If the production mechanisms are different, we have to think very hard as a field." Some, such as physicist Alex Dzierba of Indiana University, Bloomington, go even further and call the Fermilab results and an increasing number of other nonsightings from laboratories in the States and in Europe "an overwhelming body of negative evidence."

How, then, to account for the dozen sightings at different labs? There are a number of mechanisms that could lead to a hump in peakless data. Robert Chrien, a physicist at Brookhaven National Laboratory in Upton, New York, says the peaks could be artifacts caused when physicists weed out their data sets to reduce statistical noise. "As soon as you make cuts, you can introduce bias," says Chrien. "We once thought we saw a peak in a gamma ray spectrum exactly at an energy predicted by one of the theories," he says of an unrelated experiment—a peak that disappeared as soon as the biasing cuts were wiped out.

There are other possibilities, too. Dzierba suspects that some of the Θ^+ sightings are due to "ghost tracks": incorrectly reconstructed particle trails that plague high-energy physics experiments. Sometimes even the most sensitive equipment will see two particles when only one exists. Dzierba says an incorrect tally of certain reactions involving the Λ(1520), pions, and protons could make a peak at "precisely the mass of the Θ^+." Other exotic effects, such as a "reflection" of another peak, might cause a miragelike hump in the data. "There's lots of possibilities," says Barnes. However, he adds, coming up with alternative explanations for the pentaquark peak "doesn't really mean anything. You can't say it doesn't exist. The issue is going to get settled with a really good experiment."

That's what the folks at JLab hope, too. "As of now, there's no clear experimental evidence for either the existence or nonexistence of the Θ^+," says Stepan Stepanyan, a physicist at JLab, who has been performing a high-statistics search for the pentaquark at JLab. Although the team has gathered roughly five times as many data as the first JLab sighting, it is not yet ready to release its analysis. If the peak disappears or stays small, then the pentaquark will almost certainly be an artifact of the analysis. If the peak gets starker—and if the team is careful about their cuts and weeds out the ghost tracks and phantom reflections—then it will likely mean that the Θ^+ is real, and that a seemingly unstable beast gets stability from an unknown mechanism.

"The stakes are high," says Stepanyan. "If the Θ^+ exists, then our naïve picture of [nuclear] structure will change." If it doesn't exist, though, then the pentaquark hunters will add their names to the rolls of those who went hunting for big game and wound up on a wild-goose chase. **–CHARLES SEIFE**

CREDIT: K. BUCKHEIT/*SCIENCE*

Faster Than a Hyena? Running May Make Humans Special

Scientists propose that hominids evolved into long-distance runners 2 million years ago to become better scavengers on the African savanna

Depending on your point of view, last week's New York City marathon was a demonstration of athletic excellence or of unparalleled masochism. But according to a report in this week's issue of *Nature*, it was also a display of a key innovation in human evolution. University of Utah biomechanics expert Dennis Bramble and Harvard physical anthropologist Daniel Lieberman argue that the human body is exquisitely adapted for endurance running. They marshal evidence that the ability to run long distances emerged 2 million years ago, possibly enabling our ancestors to become better scavengers. If the researchers are right, running goes a long way toward explaining why our bodies are so different from those of other apes.

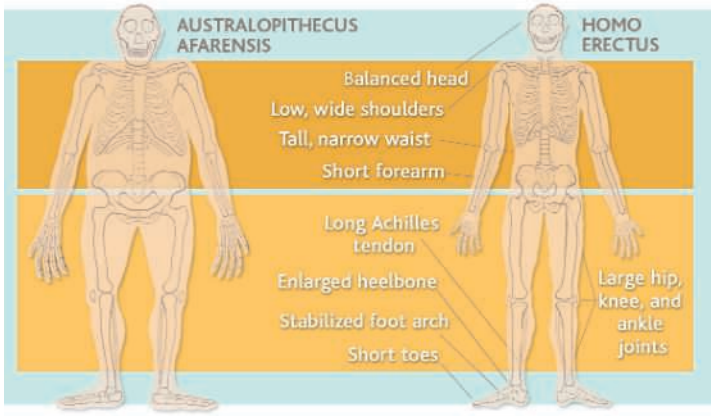
It may come as a surprise to hear that humans excel in running. Obviously, a leopard can leave us in the dust in a short sprint. But over longer distances leopards and most other mammals flag. "Most mammals can't sustain a gallop over 10 to 15 minutes," says Lieberman. Humans, on the other hand, can continue running for hours while using relatively little energy. "Humans are phenomenal endurance runners, in terms of speed, cost, and distance," says Lieberman. "You can actually outrun a pony easily." And yet, he points out, "no other primates out there endurance run."

Bramble and Lieberman believe that much of this skill comes from a large inventory of special adaptations in our muscles, tendons, and bones. They emphasize that these adaptations are not all-purpose traits that also help us walk upright. "Running is not fast walking," says Lieberman. "You do not use the same mechanics."

To identify adaptations for running, the researchers have put people and animals on treadmills and measured the activity of various muscles and ligaments, along with the forces a running body generates. The nuchal ligament, for example, which stretches from the base of the human skull to the base of the neck, stands out. "It's an elastic band that has repeatedly evolved in animals that run. Apes don't have it," says Bramble. He and Lieberman hypothesize that the nuchal ligament helps keep an endurance runner's head from bobbing violently. "Every time your heel hits the ground, your head wants to topple forward," says Lieberman.

Humans also have a special arrangement of tendons in their legs (including long Achilles tendons) that can act like springs. These tendons store about half of the energy of each stride and release it in the following one. "Chimps don't have these springs," Lieberman says, also noting that recovering energy is important for endurance running but not for sprinting.

Bramble and Lieberman have also zeroed in on the importance of a large rear end. By attaching electrodes to the gluteus maximus muscles of very cooperative volunteers, they have found that these muscles contract during each running stride, but not during walking—probably to stabilize the trunk. Chimps, by contrast, "have tiny rear ends," says Lieberman.



Endurance. Humans are unique among primates in their ability to run long distances, thanks to a number of recently identified anatomical traits.

The fossil record suggests that these adaptations for endurance running emerged together about 2 million years ago, in the early species of our own genus, *Homo*. Paleoanthropologists have long noted that some early *Homo* were markedly different from earlier hominids. *Australopithecus afarensis*, which lived from about 4 million to 3 million years ago, stood 0.9 to 1.2 meters tall and had long arms and a wide pelvis. *Homo ergaster*, which lived in Africa between 1.9 million and 1.6 million years ago, was about as tall as modern humans and had long legs and relatively short arms. Most researchers ascribed these changes to adaptations for efficient walking. "That's the standard story you'll get in most textbooks," says Lieberman. But Bramble and Lieberman have a different theory. Endurance "running is the only known behavior that would account for the different body plans in *Homo* as opposed to apes or australopithecines," says Bramble.

John Fleagle, an anatomist at Stony Brook University in New York, is impressed by Bramble and Lieberman's argument and wonders why no one thought of it before. "It's a real head-slapper," he says. A number of their predictions remain to be tested, he points out, because the fossil record of early *Homo* is still incomplete. But he expects Bramble and Lieberman's paper to generate a lot of new research.

The "sketchiest part" of their hypothesis, admits Bramble, is why hominids ran long distances. Paleoanthropologists generally agree that early *Homo* were primarily scavengers, using stone tools to cut meat off carcasses and crack open bones. "If you get [to the carcass] before the hyenas and the other hominids, you would have a lot of protein and fat at your disposal," says Lieberman.

By allowing hominids to get to more protein and fat, Lieberman suggests, running might have fueled the evolution of big hominid brains. "Large brains occur after the evolution of this modern, humanlike body form, and it may have been that the ability to do endurance running released a constraint on human evolution," he says. He speculates that the importance of endurance running only faded once humans invented hunting weapons.

"The importance of running to the tens of thousands of people who run marathons every year is not just a fluke," says Lieberman. "I think it's a result of some important evolutionary history. It may have been lost with the invention of the bow and arrow, but the traces are still there in our bodies."

–CARL ZIMMER

Carl Zimmer is the author of *Soul Made Flesh: The Discovery of the Brain—and How it Changed the World*.

CREDITS (TOP TO BOTTOM): MARY ALTAFFER/ASSOCIATED PRESS; ADAPTED FROM LASZLO MESZOLY

Published by AAAS

Of Worms, Mice, and Very Old Men and Women

TORONTO, CANADA—More than 5000 experts met here from 26 to 30 October for the annual meeting of the American Society of Human Genetics. Longevity, milk digestion, and cancer were among the topics.

If your grandparents all lived to a ripe old age, you probably hope that "good genes" will bring you long life as well. Researchers in New York City have been exploring that notion by studying the physiology and DNA of Ashkenazi Jews who have lived for almost a century or longer. Last year, the scientists found that these elders have an unusual profile of lipids in their blood that may explain their longevity. In Toronto, the same group reported that their subjects tend to have a gene variant involved in a lipid pathway previously tied to long life in *Caenorhabditis elegans*, the model worm.

This may be the first evidence that a similar longevity gene pathway acts in both worms and people, says Cynthia Kenyon, a researcher on aging at the University of California, San Francisco, who finds the link "really interesting." In another talk, collaborators of the New York group reported that the centenarian Ashkenazis also tend to carry a particular variant of a gene that causes early aging in mice when it's turned off.

In the genes. Some centenarians may owe their long life to mutations in genes that govern blood lipids.

The Longevity Genes Project led by Nir Barzilai of Albert Einstein School of Medicine includes more than 300 Ashkenazi Jews with an average age of 98, as well as their offspring and age-matched controls to the offspring from Ashkenazi families with average life spans. The blood of these centenarians contains especially large lipoprotein particles, which are normally seen only in young, exercising adults (*Science*, 17 October 2003, p. 373). Barzilai's team also has shown that these elderly Jews and their offspring tend to have a mutation in the gene for cholesteryl ester transfer protein, which raises levels of the "good" high density lipoprotein (HDL) cholesterol and also increases particle size of both HDL and low density lipoprotein (LDL), the bad kind.

At the meeting, team member Gil Atzmon reported that another gene involved in lipid metabolism appears to protect the Jewish centenarians. The gene codes for a protein called apolipoprotein CIII (ApoC-III), which is a component of LDL. ApoC-III stimulates production of harmful lipids called triglycerides; high levels of these compounds raise the risk of cardiovascular disease.

The centenarian families were more likely to have two copies of an *ApoC-III* gene with a particular one-base change in its promoter region than were controls (25% compared to 11%). Those with two copies of the *ApoC-III* variant had lower blood levels of the protein, larger lipid particles, and lower triglyceride levels, as well as lower rates of hypertension and resistance to insulin, two common signs of aging. The "most dramatic" result, says Barzilai, emerged when his team looked for this mutation in blood samples from a separate group of Ashkenazis, some of whom had died: Those with two copies of the *ApoC-III* gene variant tended to live 4 years longer, on average.

Intriguingly, the *ApoC-III* gene is controlled by the gene for a transcription factor called FOXO1. The worm version of this gene is involved in a regulatory pathway that governs aging; altering the pathway can extend, even double, a worm's life span. In the worm, this pathway turns on a set of genes that include those producing lipids, Kenyon's group recently found. That suggests the same set of genes determines life span in both people and worms, Kenyon says: "I'm not sure anyone else has made this connection."

In collaboration with Barzilai's group, Harry Dietz's team at Johns Hopkins University has also examined the centenarian Ashkenazi Jews' blood samples for variants of *KLOTHO*, a gene that when deactivated in mice leads to what looks like early aging. In 2002, Dietz's team reported on a study of the gene in a group of elderly Czechs: Having just one copy of a certain variant of *KLOTHO* seemed to help people live longer, whereas two copies of this variant led to earlier death.

At the meeting, postdoc Dan Arking reported that he and his Hopkins colleagues have found similar results in the centenarian Ashkenazi Jews, 44% of whom have died since the study began in 1998. Compared with the study subjects with one copy of the *KLOTHO* gene variant, those with no copies were twice as likely to have died; those with two copies had a 4.5-fold higher risk. Not only did those with one copy of the allele live longer, but they also had lower blood pressure and higher HDL levels, which means a lower risk of stroke, says Arking.

The quest for longevity genes "is really tough" because life span is almost certainly controlled by a multitude of genes, says George Martin of the University of Washington, Seattle, editor-in-chief of the Science of Aging Knowledge Environment. Originally skeptical that any single mutation could be tied to human longevity, Martin now says that Barzilai's team "is making progress" at proving him wrong.

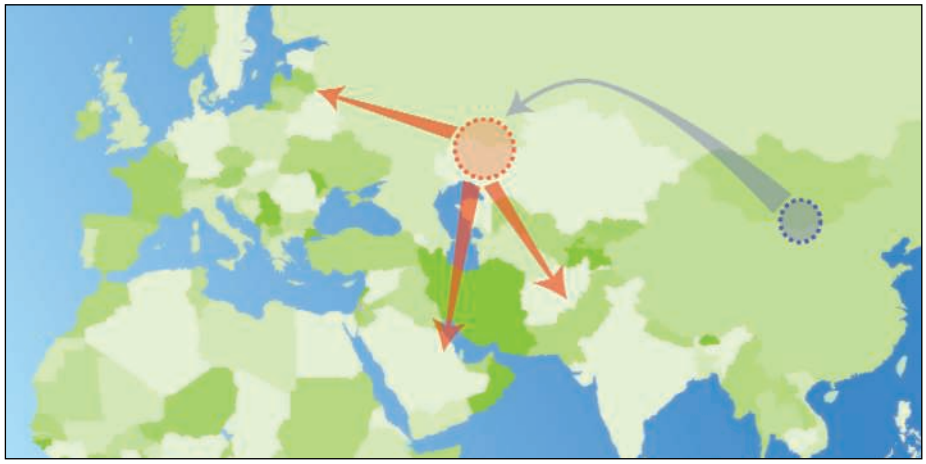
Ural Farmers Got Milk Gene First?

By some estimates, less than half of all adults can easily digest milk, a trait believed to have first appeared in people who kept dairy animals. Now scientists have traced the genetic roots of milk tolerance to the Ural mountains of western Russia, well north of where pastoralism is thought to have begun. The surprising result may support a theory that nomads from the Urals were one of two major farmer groups that spread into Europe, bringing the Indo-European languages that eventually diverged into the world's largest family of modern languages.

Almost all mammalian babies produce lactase, the enzyme that digests the milk sugar lactose. But in most animals and many people, the lactase gene is gradually turned off after infancy, leaving them unable to tolerate milk as adults. Two years ago, a team led by Leena

CREDIT: RON WINN/ASSOCIATED PRESS

Published by AAAS



Milk route. A new study suggests that tribes from the Asian steppes (blue circle) migrated to the Ural mountains, where they mixed with locals (red circle), generating a gene variant endowing lactose tolerance that Ural farmers later spread.

Peltonen of the University of Helsinki, Finland, and the University of California, Los Angeles, identified mutations near the lactase gene that are associated with adult lactose tolerance and likely play a role in regulating the lactase gene. Now, Peltonen's team has tried to trace the origins of lactose tolerance by looking at 1611 DNA samples from 37 populations on four continents.

The populations having the greatest DNA sequence diversity around the lactase gene mutations—suggesting that lactose tolerance first appeared in them—include the Udmurts, Mokshas, Ezras, and other groups that originally lived between the Ural mountains and the Volga River. The trait most likely developed 4800 to 6600 years ago, Peltonen says. Her team linked the lactase gene changes to an ancestral variant that these groups apparently got from intermixing with tribes migrating from the Asian steppes.

After the Ural peoples gained this earlier form of the lactase gene, the lactose tolerance mutation "probably emerged by chance," says Peltonen, and then remained because it was beneficial for milk consumption. The Ural groups then likely later spread the variant to Europe—especially northern Europe, which has the highest lactose tolerance today—and the Middle East. The findings support the somewhat controversial theory that nomadic herders known as Kurgans expanded into Europe from the southern Urals 4500 to 3500 years ago, bringing Indo-European languages with them, according to Peltonen.

"I find [the new study] very interesting," says population geneticist Luigi Luca Cavalli-Sforza of Stanford University. He notes that a competing idea for explaining the origin of the Proto-Indo-Europeans is that they were crop-growing farmers from the Anatolia region in modern Turkey (*Science*, 27 February, p. 1323). But the milk study reinforces Cavalli-Sforza's view that both theories are correct: Indo-Europeans migrated to Europe in two waves, first from Turkey and later from the Urals.

Other geneticists caution that trying to pin down where a gene variant originated is tricky because the people in whom it's most common today may have migrated from somewhere else, or the original population could now be extinct. But if the milk gene's origin holds up, linguists and archaeologists will have new food for thought.

CREDITS (TOP TO BOTTOM): SOURCE: N. ENATTAH; J. A. MARTIGNETTI

New Prostate Cancer Genetic Link

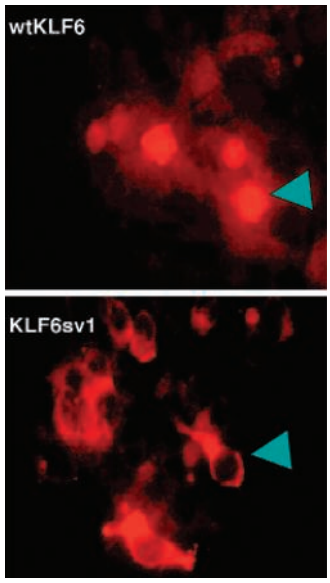
Prostate cancer strikes one in six men on average, and many researchers are looking for the genetic factors that contribute to an individual's risk. In Toronto, cancer geneticists presented data indicating that a relatively common variant of a gene involved in cell growth can raise a man's prostate cancer risk. The researchers also suggested how this variant may spur cancerous growth.

Researchers at Mount Sinai School of Medicine in New York City had already found mutated versions of the gene, Kruppel-like factor 6 (*KLF6*), in many prostate tumors; the mutations disrupt *KLF6*'s normal role of inhibiting cell growth (*Science*, 21 December 2001, p. 2563). Those mutations may have arisen late in life, but could more subtle variations in the gene, ones present from birth, set the stage for prostate cancer down the road? The same Mount Sinai team, led by John Martignetti, and collaborators have now examined this question by drawing on registries from three major cancer centers. The investigators screened for a previously identified *KLF6* variant, a single-base mutation, in blood samples from 3411 prostate cancer patients—some with a family history of prostate cancer, some without—and controls.

The variant was overrepresented in the patients, suggesting that it may predispose men to prostate cancer. About 17% of the patients with a family history of the disease and 15% of the patients with no such history carried at least one copy of the variant, whereas only 11% of the controls possessed a copy. From that data, the researchers calculate that, compared with men lacking the variant, men with at least one copy have an increased prostate cancer risk—about a 50% hike.

The team also investigated how this particular variation in the *KLF6* gene changes its function. They found that the protein encoded by *KLF6* can come in three forms of differing sizes; cells with the variant make more of the two truncated forms. Instead of entering the cell nucleus and suppressing cell growth, these shortened KLF6 proteins stay in the cytoplasm, where they have the opposite effect. Tipping the balance of KLF6 proteins toward the short versions could promote cell growth and therefore explain why the *KLF6* variant raises cancer risk, Martignetti says.

Variations in at least three other genes have been identified as raising a man's prostate cancer risk, although the links haven't always been confirmed in different populations. To some cancer researchers, Martignetti's team has made a convincing case about *KLF6*. "It's about as solid as it could be" for an initial study, says Sean Tavtigian of the International Agency for Research on Cancer in Lyon, France. And although the *KLF6* variant alone may only slightly raise a man's prostate cancer risk, it could act in concert with other gene variants, Tavtigian notes. If researchers can pin down how those risks add up, clinicians might one day screen a man's DNA to determine his true risk of prostate cancer—and prescribe preventive strategies to especially susceptible men.



Off balance. The protein encoded by *KLF6* normally appears in the nucleus (*top*), but mutations in the gene lead to more of it in the cytoplasm.

—JOCELYN KAISER

Published by AAAS

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

Assisting, But Not Dictating

WHEN READING A JOURNAL SUCH AS *SCIENCE*, one is easily seduced into believing that empirical evidence can resolve moral disputes. In his Letter "Human being redux" (16 Apr., p. 388), M. S. Gazzaniga defends human embryonic stem cell research because of the vast discrepancy between a tiny ball of cells that can fit on the head of a pin and a live human being. J. T. Durkin ("The case against stem cell research," Letters, 3 Sept., p. 1402) minimizes this disparity by emphasizing that "[t]he embryo and the adult are different stages in the development of the human being." By referring to empirical information, they seem to think that the right (good) social policy for stem cell research can be justified. G. E. Moore's philosophical position, known as the naturalistic fallacy, argues that "goodness" is indefinable, and therefore its meaning cannot be logically derived by empirical means (*1*). That is, our biological underpinnings cannot prescribe what is good and right. However, facts in combination with a democratic ethic can assist in determining a policy decision. Although individuals will differ in their opinions, a democracy can decide whether the benefits of embryonic stem cell research outweigh any disadvantages. Science can assist in making this decision, but cannot dictate it (*2*).

HOWARD H. KENDLER

Department of Psychology, University of California, Santa Barbara, Santa Barbara, CA 93106, USA.

References

1. G. E. Moore, *Prinicipia Ethica* (Cambridge Univ. Press, Cambridge, 1903).
2. H. H. Kendler, *Amoral Thoughts About Morality: The Intersection of Science, Psychology, and Ethics* (Charles C. Thomas, Springfield, IL, 2000).

Microbial Life in the Atacama Desert

IN THEIR REPORT "MARS-LIKE SOILS IN THE Atacama Desert, Chile, and the dry limit of microbial life," R. Navarro-Gonzáles *et al.* found only very low levels of culturable bacteria in the Mars-like soils of the Atacama Desert, and they did not recover DNA (Reports, 7 Nov. 2003, p. 1018). In contrast, we have found easily cultured, low numbers of bacteria and recoverable bacterial DNA from soils in the extreme arid core of the Atacama Desert in northern Chile.

Soil samples taken from a 4500-m elevational transect just south of the Tropic of Capricorn (–24ºS) all yielded culturable bacteria on R2A agar (*1*, *2*), including samples from elevations of absolute desert that have not harbored plant life for a million years or more. Four of our samples were taken in the vicinity of the dry Yungay region, in close proximity to those studied by Navarro-Gonzáles *et al.* (elevation ~1000 m: S 24º4.16', W 69º51.98' and S 24º4.185', W 69º51.968'). Our three closest sites (987 m: S 24º4.517', W 70º12.555'; 1315 m: S 24º21.787', W 69º56.757'; and 1931 m: S 24º28.135', W 69º24.472') yielded counts of 1.3×10^5, 5.4×10^3, and 9.1×10^4 CFU/g of dry soil, respectively. A fourth site (703 m: S 23º57.417', W 70º17.157') yielded only 1 or 2 colonies per plate, which is a value too close to the detection limit of the spread plating method to quantify accurately but is still higher than that reported by Navarro-Gonzáles *et al.* (<10 colonies found on 100 plates).

A rock formation in the extremely arid Atacama Desert in northern Chile.

Bacterial DNA was successfully extracted (*3*) from all of our samples (Navarro-Gonzáles *et al.* report no recovery of DNA from the Yungay samples), and 16S rRNA genes were amplified (*4*, *5*) and profiled by denaturing gradient gel electrophoresis (DGGE). Statistical analysis of DGGE profiles demonstrates a similar bacterial community structure in samples taken from soil profiles in the absolute desert portions of our Atacama transect. This community structure is quite different from that found in profiles from vegetated zones supported by fog or precipitation below (<500 m) and above (>2500 m) the absolute desert, respectively. Our results demonstrate the existence of life in one of the driest regions on Earth. We may have been able to demonstrate life because we sampled at a depth of 20 to 30 cm, in comparison to Navarro-Gonzáles *et al.*, who sampled the upper 10 cm of the soil. This only emphasizes the critical nature of the sampling protocol used in any extreme environment on Earth and particularly on Mars.

R. M. MAIER,[1] K. P. DREES,[4] J. W. NEILSON,[1] D. A. HENDERSON,[2] J. QUADE,[3] J. L. BETANCOURT[5]

[1]Department of Soil, Water and Environmental Science, [2]Department of Animal Sciences, Division of Epidemiology/Biostatistics, [3]Department of Geosciences, University of Arizona, Tucson, AZ 85721, USA. [4]Department of Civil Engineering, University of Minnesota, Minneapolis, MN 55455, USA. [5]U.S. Geological Survey, 1675 West Anklam Road, Tucson, AZ 85745, USA.

References

1. *Difco Manual* (Difco Laboratories, Spark, MD, ed. 11, 1998), p. 421.
2. K. P. Drees, Ph.D. dissertation, University of Arizona, Tucson, AZ (2004).
3. Fast DNA Spin Kit for Soil, Qbiogene, Carlsbad, CA.
4. G. M. Colores, R. E. Macur, D. M. Ward, W. P. Inskeep, *Appl. Environ. Microbiol.* **66**, 2959 (2000).
5. M. J. Ferris, G. Muyzer, D. M. Ward, *Appl. Environ. Microbiol.* **62**, 340 (1996).

Response

IN OUR PAPER, WE REPORTED EXTREMELY LOW levels of culturable organisms and no recoverable DNA in the surface soils of the extreme arid core of the Atacama Desert near the abandoned town of Yungay. We could not claim that there was no life in these soils on the basis of our results, and therefore we presented our data as indicating an upper limit of 100 culturable heterotrophic bacteria per gram of soil (see fig. 2E of our Report) for surface materials. This upper limit is orders of magnitude less than the concentrations of bacteria found in soils south of this Mars-like region of the Atacama. In more recent published work (*1*), we have reported that below the surface, there are discrete layers with higher numbers of culturable bacteria. For example, at a Yungay site, we have found negligible levels of bacteria at the surface (<100 CFU/g) but recovered less than 1×10^2 to 2.96×10^5 CFU/gram of soil in subsurface layers (*1*). In addition, we have conducted an extensive survey of surface and subsurface soils in the arid core of the Atacama (*1*–*4*). The data presented by Maier *et al.* for subsurface samples are consistent with our published work [our Report; (*1*–*4*)] and do not necessitate any reassessment or reevaluation of the conclusions of our Report. We agree with their conclusion regarding the critical nature of the sampling protocol used in any extreme environment on Earth and Mars.

RAFAEL NAVARRO-GONZÁLEZ,[1] FRED A. RAINEY,[2] CHRISTOPHER P. MCKAY[3]

[1]Laboratorio de Química de Plasmas y Estudios

Planetarios, Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Circuito Exterior, Ciudad Universitaria, Apartado Postal 70-543, México D.F. 04510, Mexico. E-mail: navarro@nuclecu.unam.mx. [2]Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA. E-mail: frainey@lsu.edu. [3]Space Science Division, NASA-Ames Research Center, Moffett Field, CA 94035–1000, USA. E-mail: cmckay@mail.arc.nasa.gov

References
1. D. R. Bagaley *et al.*, *Int. J. Astrobiol.*, suppl. 1, p. 78 (2004) (Abstracts from the Astrobiology Science Conference 2004).
2. D. R. Bagaley, R. Navarro-González, B. Gomez-Silva, C. P. McKay, F. A. Rainey, abstract presented at the 104th General Meeting of the American Society of Microbiology, New Orleans, LA, 2004.
3. R. Navarro-González *et al.*, *Int. J. Astrobiol.*, suppl. 1, p. 82 (2004) (Abstracts from the Astrobiology Science Conference, 2004).
4. R. Navarro-Gonzalez *et al.*, abstract presented at the 35th Plenary Meeting of the Committee on Space Research, Paris, 18 to 25 July 2004.

Varshavsky's Contributions

WE ARE WRITING TO EXPRESS OUR enthusiasm that the discovery of the ubiquitin conjugation system has been acknowledged with the award of the Nobel Prize in Chemistry to three outstanding biochemists: Avram Hershko and Aaron Ciechanover of the Technion Institute in Israel, and Irwin Rose of the University of California at Irvine ("Gold medal from cellular trash," G. Vogel, News Focus, 15 Oct., p. 400). Unraveling the chemistry that underlies the attachment of ubiquitin to proteins that are destined to be degraded was a magnificent achievement and is fully deserving of this recognition.

The mechanism of ubiquitin conjugation and its role in proteolysis was selected for recognition by the Nobel Committee in large part because of the vital role that the ubiquitin-proteasome system (UPS) plays in the physiology of cells and organisms. Investigations on the physiological functions of the UPS, which dominate current research in this field, were pioneered largely by Alexander Varshavsky of the California Institute of Technology. Several core principles that guide our current understanding of the ubiquitin system had their origins in Varshavsky's work, including the following: (i) the UPS is the predominant mechanism for selective protein turnover in the cytoplasm and is essential for cellular function; (ii) in addition to its role in turning over damaged proteins, the UPS controls diverse physiological processes such as the cell cycle, DNA repair, and stress responses; (iii) ubiquitin ligases (E3s) are highly specific receptors that underlie the remarkable specificity of ubiquitination by binding to defined sequences within proteins (degrons); and (iv) a ubiquitin chain linked via the lysine-48 residue of ubiquitin governs targeting of substrates to the proteasome for degradation. In addition to these fundamental contributions, Varshavsky ushered the ubiquitin field into the age of molecular genetics by identifying mutants and characterizing

"We suggest that the impact of Varshavsky's work on the physiology of the ubiquitin system and its relationship to fundamental processes such as mitosis and chromosome segregation justifies serious consideration for a future Nobel Prize in Physiology or Medicine."

–*BAUMEISTER* ET AL.

the genes that define ubiquitin and the E1, E2, and E3 components of the conjugation cascade. In our opinion, the appreciation of the significance of the UPS owes more to Varshavsky's work than to that of any other individual.

We suggest that the impact of Varshavsky's work on the physiology of the ubiquitin system and its relationship to fundamental processes such as mitosis and chromosome segregation justifies serious consideration for a future Nobel Prize in Physiology or Medicine. As we extend our heartfelt congratulations to the winners of this year's Nobel Prize in Chemistry, we wish to make it clear that Varshavsky's contributions are also deeply respected by his colleagues.

WOLFGANG BAUMEISTER,[1] ANDREAS BACHMAIR,[2] VINCENT CHAU,[3] ROBERT COHEN,[4] PHIL COFFINO,[5] GEORGE DEMARTINO,[6] RAYMOND DESHAIES,[7] JUERGEN DOHMEN,[8] SCOTT EMR,[9] DANIEL FINLEY,[10] RANDY HAMPTON,[9] CHRISTOPHER HILL,[11] MARK HOCHSTRASSER,[12] ROBERT HUBER,[1] PETER JACKSON,[13] STEFAN JENTSCH,[1] ERICA JOHNSON,[14] YONG TAE KWON,[15] MICHELE PAGANO,[16] CECILE PICKART,[17] MARTIN RECHSTEINER,[11] MARTIN SCHEFFNER,[18] THOMAS SOMMER,[19] WILLIAM TANSEY,[20] MIKE TYERS,[21] RICHARD VIERSTRA,[22] ALLAN WEISSMAN,[23] KEITH D. WILKINSON,[24] DIETER WOLF[25]

[1]Max Planck Institute of Biochemistry, Martinsried D-82152, Germany. [2]Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany. [3]Pennsylvania State University at Hershey, Hershey, PA 17033, USA. [4]University of Iowa, Iowa City, IA 52242, USA. [5]University of California at San Francisco, San Francisco, CA 94143, USA. [6]University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [7]California Institute of Technology, Pasadena, CA 91125, USA. [8]University of Cologne, Germany. [9]University of California at San Diego, La Jolla, CA 92093, USA. [10]Harvard Medical School, Boston, MA 02115, USA. [11]University of Utah, Salt Lake City, UT 84132, USA. [12]Yale University, New Haven, CT 06520, USA. [13]Stanford University Medical School, Stanford, CA 94305, USA. [14]Thomas Jefferson University, Philadelphia, PA 19107, USA. [15]University of Pittsburgh, Pittsburgh, PA 15260, USA. [16]New York University, New York, NY 10016, USA. [17]Johns Hopkins University, Baltimore, MD 21205, USA. [18]University of

TECHNICAL COMMENT ABSTRACTS

COMMENT ON "Small Bilaterian Fossils from 40 to 55 Million Years Before the Cambrian"

Stefan Bengtson, Graham Budd

The report by Chen *et al*. of coelomate bilaterian fossils from ancient phosphorites (Research Articles, 9 July 2004, p. 218) is not well founded. The morphological features reported can be simply accounted for by familiar taphonomic and diagenetic processes. The structures may well be eukaryotic microfossils, but their present appearance has little resemblance to the once-living organisms.

Full text at www.sciencemag.org/cgi/content/full/306/5700/1291a

RESPONSE TO COMMENT ON "Small Bilaterian Fossils from 40 to 55 Million Years Before the Cambrian"

Jun-Yuan Chen, Paola Oliveri, Eric Davidson, David J. Bottjer

The premise presented by Bengtson and Budd is incorrect, and their example is irrelevant. We provide two new images of the holotype specimen that demonstrate that the definitive characters of the specimen discussed by us in the original report are even more extensively evident than was initially apparent.

Full text at www.sciencemag.org/cgi/content/full/306/5700/1291b

Letters

Konstanz, 78457 Konstanz, Germany. [19]Max Delbrueck Centre for Biomedical Research, Berlin 13122, Germany. [20]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [21]Samuel Lunenfeld Research Institute, Toronto, ON MG5 1X5, Canada. [22]University of Wisconsin, Madison, WI 53706, USA. [23]National Cancer Institute, National Institutes of Health, Frederick, MD 21702, USA. [24]Emory University, Atlanta, GA 30322, USA. [25]University of Stuttgart, D-70569 Stuttgart, Germany.

CORRECTIONS AND CLARIFICATIONS

Reports: "Mitotic Golgi partitioning is driven by the membrane-fissioning protein CtBP3/BARS" by C. Hidalgo Carcedo *et al.* (2 July, p. 93). This paper reported that BARS is crucially involved in mitotic Golgi partitioning and entry into mitosis. CtBP3/BARS (BARS) is a protein involved in Golgi membrane fission [S. Spanò *et al.*, *J. Biol. Chem.* **274**, 17705 (1999); R. Weigert *et al.*, *Nature* **402**, 429 (1999)] and a member of the CtBP family that comprises CtBP1 and CtBP2, both of which are transcriptional co-repressors [G. Chinnadurai, *Mol. Cell* **9**, 213 (2002); G. Chinnadurai, *Bioessays*. **25**, 9 (2003)]. BARS is almost certainly a *ctbp1* gene product and therefore a splice variant of CtBP1. A KO mouse has been generated in which both *ctbp1* and *ctbp2* have been deleted (and which therefore also lacks BARS). This KO is embryonically lethal, but cells derived from these embryos proliferate normally, indicating that the partitioning of their Golgi complex should occur during mitosis (although it is not clear that the Golgi partitions normally) [J. D. Hildebrand, P. Soriano, *Mol. Cell. Biol.* **22**, 5296 (2002)]. This is apparently discrepant with the report by Hidalgo Carcedo *et al.* that BARS is crucially involved in mitotic Golgi partitioning and entry into mitosis. Similar discrepancies between KO and classical cell biological studies in cultured cells are frequent and provide useful insights into the process under study [M. Pagano, P. K. Jackson, *Cell* **118**, 535 (2004)]. The following is a brief discussion of a few hypotheses that can cast light on this specific case. First, more than one fission mechanism might be involved in mitotic Golgi partitioning. For instance, mitotic Golgi fragmentation involves two stages, one consisting of the consumption of Golgi membranes by the irreversible budding of COPI vesicles, and the second, of the tubulation of Golgi cisternae followed by their cleavage into smaller pieces [J. Shorter, G. Warren, *Annu. Rev. Cell. Dev. Biol.* **18**, 379 (2002)]. The latter component is likely to be the one that is dependent on BARS. It is possible that in embryonic cells lacking BARS, the COPI-dependent mechanism might carry the Golgi partitioning process far enough to allow mitosis to proceed. Another possibility is that in these cells, once the Golgi cisternae have been transformed into tubules during mitosis (presumably via phosphorylation of the relevant golgins) (Shorter and Warren), a dynamin-like protein is able to cleave these tubes into small pieces. Second, BARS might not be a core fission protein, but rather a regulator, which could be replaced in BARS-null cells by a related gene with a similar function. Finally, it is possible that the Golgi structure in embryonic cells is organized differently and does not require the BARS-controlled machinery to enter mitosis. This last possibility is supported by morphological studies that are presently in progress. The above mechanisms (and possibly others) might allow the cells to undergo mitosis and execute Golgi partitioning even in the absence of BARS. This could result in mitotic Golgi phenotypes that might or might not be different from those in control cells.

Comment on "Small Bilaterian Fossils from 40 to 55 Million Years Before the Cambrian"

Chen *et al.* (*1*) reported coelomate bilaterians from the ~600-million-year-old Doushantuo phosphorites in southern China. Such a find might meet some common expectations of small, simple bilaterians emerging after the worldwide glaciations of the Neoproterozoic. The interpretation is not well founded, however, because it fails to take into full account taphonomy (changes in the organism after death) and diagenesis (changes in the sediment after deposition).

The specimens presented by Chen *et al.* (*1*) represent a common mode of preservation of microfossils in phosphatic sediments, including those of the Doushantuo (*2*) and the overlying Dengying (*3*) formations: a more or less undeformed outer membrane; a shrunken, irregular internal mass, often connected to the outer membrane by occasional threads or sheets; and thin layers of diagenetic mineral, commonly apatite (calcium phosphate), lining the surfaces of the resulting cavities (Fig. 1). Such diagenetic minerals usually have a characteristic crystallographic structure due to the growth direction normal to the encrusted surface. Chen *et al.* have provided no information on the structure of the layers they interpret as cellular, but even the published figures show clear evidence of diagenetic origin:

(i) The layers have a regular banding of color and thickness that is different between the specimens but consistent within the individual specimens, whether counted from the outer wall inward or from the central body outward. In the direction toward what they describe as the coelomic lumen, Chen *et al.* showed a thickness sequence of approximately 2 + 2 + 5 μm [figure 1A in (*1*)]; one of 3 + 5 + 5 μm, with the first layer considerably darker than the subsequent two [figure 1B in (*1*)]; and one of 2 + 5 μm [figure 1C in (*1*)]. This pattern defies biological explanation but is easily explained as representing two to three generations of diagenetic overgrowth (Fig. 1, C to E).

(ii) Rather than being sinuously folded, as would be expected from deformed tissue layers, the layers consistently have their convex features directed toward the putative coelomic lumen. This is a typical feature of diagenetic crusts, in which irregularities on the overgrown surface serve as nuclei for spherulitic fans (Fig. 1, B to E).

(iii) The layers show typical cavity-filling geometry. The outermost layer is missing in narrow spaces where earlier growth left no room for it [compare, for example, figure 1B-1, lower part, in (*1*) and Fig. 1C, arrow, in this comment].

(iv) The layers show conspicuous dark lines perpendicular to the surfaces. Chen *et al.* (*1*) refer to these as "cellular structure preserved." Cell walls may indeed be preserved in these sequences (*3*–*5*), typically by internal encrustation by diagenetic minerals (*6*). Surface-perpendicular lines within diagenetic crusts, however, more likely represent fine cracks propagated along the directions of surface-normal acicular crystallites. The fact that the lines in the specimens depicted by Chen *et al.* [for example, figure 2, C and D, in (*1*)] commonly continue across the boundaries between the layers is strong evidence that they are propagating cracks rather than cell boundaries.

Chen *et al.* concede that one thin partial coating of the layers is of diagenetic origin. They do not address, however, any of the clear indications of diagenesis in the main layers, but simply assert that they represent cellular layers because they have been consistently observed in independent specimens of the same morphological organization and similar dimensions. In view of the fact that the supposedly bilaterian specimens have been selected from among 50,000 to 100,000 microfossils, the perceived consistencies are far from impressive. The supposed mouth and anus are reported from two specimens each. The putative pharynx, gut, and coeloms are reported from all 10 selected specimens, but these features will be present by construct if specimens are selected that happen to have the internal lump of

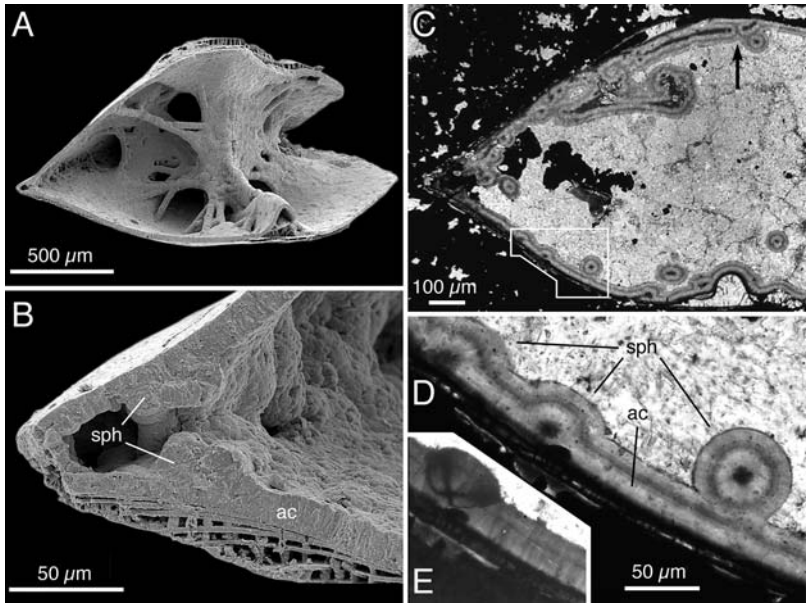


Fig. 1. Lingulate brachiopod *Linnarssonia* from Middle Cambrian phosphatized limestone in south Sweden (locality Kvasa, Scania). (**A** and **B**) Scanning electron microscopy pictures of broken specimen. Swedish Museum of Natural History (SMNH) Br138245. Posterior parts of dorsal (bottom) and ventral (top) valves preserved, together with shrunken soft tissues still connected by stretched threads or sheets to shell, mainly to areas of strong epithelial adherence (sites of muscle attachment). All internal cavities are lined by diagenetic apatite consisting of surface-normal acicular crystallites (ac), in places forming spherulitic fans (sph). (**C** to **E**) Thin section of similarly preserved specimen in limestone matrix. SMNH X1520 [also shown in figure 3, D and E, in (*7*)]. Plane-polarized light (C and D) and crossed nicols (E); frame in (C) indicates position of (D). Note layered structure of diagenetic lining, ordered crystallography [see polarizing cross in spherulite in (E)], inwardly convex spherulitic fans growing over irregularities on encrusted surface, absence of later-formed layer in narrow spaces [arrow in (C)], and surface-normal lineations (corresponding to direction of acicular apatite crystallites) crossing boundary between generations of diagenetic lining. This kind of diagenetic lining, also exemplified in (*2*), (*3*), and (*6*), is analogous to the purported cellular layers of the Doushantuo bilaterians as interpreted by Chen *et al.* (*1*).

shrunken matter touching or connected with the outer wall in two places in the plane of section. No information is given, other than conjecture, on the three-dimensional structure of the figured specimens, and there is no account of the variability of any of the 50,000 to 100,000 microfossils not selected as bilaterians. As a result, even a reader unacquainted with diagenesis would be hard put to identify any morphological regularities in the specimens.

Similarly, the "regularly spaced pits" on the outer surface, interpreted as sensory organs, appear to represent irregularities formed by mineral growths penetrating the surface of the original microfossil and accentuated by typical spherulite fans in the cavity-lining layers (see, for example Fig. 1D, upper left, and E, in this comment). Numerous such structures exist in the specimens depicted by Chen *et al*.—for example one specimen [figure 1B in (*1*)] shows at least a dozen of them along the periphery and half that number around the area that Chen *et al*. interpret as the gut—and the ones selected as pits differ in no important way from the others. Even the designated pits are not regularly spaced as claimed, as can be easily seen from the figured specimens.

When taphonomy and diagenesis are taken into account, the evidence that these fossils preserve minute coelomate bilaterians disappears. The objects illustrated and described by Chen *et al*. (*1*) may well be eukaryotic microfossils, but their reconstructed morphology as bilaterians is an artifact generated by cavities being lined by diagenetic crusts. The appearance of the fossils now has little resemblance to that of the living organisms that generated them.

To paraphrase Theodosius Dobzhansky: Nothing in paleontology makes sense except in the light of taphonomy and diagenesis.

Stefan Bengtson*
Department of Palaeozoology
Swedish Museum of Natural History
Box 50007
SE-104 05 Stockholm, Sweden

Graham Budd
Department of Earth Sciences,
Palaeobiology
Norbyvägen 22
SE-752 36 Uppsala, Sweden

**To whom correspondence should be addressed:*
E-mail: stefan.bengtson@nrm.se

References

1. J.-Y. Chen *et al.*, *Science* **305**, 218 (2004); published online 3 June 2004 (10.1126/science.1099213).
2. S. Xiao, X. Yuan, A. H. Knoll, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13684 (2000).
3. Z. Yue, S. Bengtson, *Lethaia* **32**, 181 (1999).
4. S. Xiao, Y. Zhang, A. Knoll, *Nature* **391**, 553 (1998).
5. S. Bengtson, Z. Yue, *Science* **277**, 1645 (1997).
6. S. Bengtson, in *The New Panorama of Animal Evolution, Proceedings of the 18th International Congress of Zoology*, A. Legakis, S. Sfenthourakis, R. Polymeni, M. Thessalou-Legaki, Eds. (Pensoft, Sofia, Bulgaria, 2003), pp. 289–300.
7. S. Bengtson, *Lethaia* **9**, 185 (1976).

10 June 2004; accepted 16 October 2004

Response to Comment on "Small Bilaterian Fossils from 40 to 55 Million Years Before the Cambrian"

The comment of Bengtson and Budd (*1*) is predicated on a preconception that any structures in sectioned Doushantuo microfossils that are claimed to represent cellular features in the original animal must instead be diagenetic artifacts. This preconception is demonstrably false. There are many examples of cellular structures in sections through other Doushantuo microfossils (Fig. 1, A to D, and F) preserved in phosphorite, as are the *Vernanimalcula* sections (*2*). Indeed, a scanning electron microscopy (SEM) image (Fig. 1E) displays a very similar cleavage form, as does the section in Fig. 1D. Both large and small blastomeres are demonstrated in both images. In the sections, the locations of at least two essentially definitive cellular features of the original embryonic cells, vis, their boundaries [(that is, cell walls) (Fig. 1, A to D and F)] and their nuclei (Fig. 1A) can easily be seen. Thus, it is counterfactual to deny preservation of structural morphology at the cellular level in this kind of material.

Turning now to *Vernanimalcula*, Bengston and Budd claim that the putative cellular structures [table 1 in (*2*)] are merely cracks in the fossil because they extend across to adjacent layers. This argument is false. The features in question here are the regularly spaced crosswise seams visible in many of the morphological layers of the holotype fossil, because these are in the positions expected of cell boundaries. Perhaps the image used by Bengtson and Budd was of insufficient resolution to reveal the details adequately; here, we offer another view (Fig. 1G), taken with polarized light under crossed nicols. There are indeed some true cracks that traverse the holotype fossil at the plane of focus shown. However, a careful count of all the crosswise partitions or seams tracing from the mesodermal to adjacent ectodermal or endodermal layers shows that only 17 out of 83 could possibly be accounted for as cracks using the criterion of Bengtson and Budd, that the seam is not confined to a single morphological layer. The large majority of the crosswise seams are indeed best taken as the remains of cell boundaries, although this is not a point we made in (*2*). Furthermore, Fig. 1G shows another prominent and revealing feature that directly affects this argument: Virtually every one of the cuboidal areas delimited by the periodic seams has a greenish spot of birefringence within it, usually toward the middle. The proposition that the seams are diagenetic cracks provides no explanation whatsoever for the striking periodicity of these spots. However, the proposition that the seams delimit the remains of cells provides an excellent explanation for their periodicity: The greenish spots could represent a systematic compositional feature caused during phosphatization by the remains of coagulated constituents of each cell, cell by cell, or they could be the result of mineral accumulation on what were the cell nuclei. Note that in what is easily recognizable as a whitish diagenetic deposit at the posterior end of the fossil there are also some greenish spots of birefringence, but they are much coarser, no two are the same, and they display no order or periodicity.

Images of the holotype specimen at a deeper plane of focus (Fig. 2) add two new items of information. First, and most important for the arguments raised by Bengtson and Budd, the "supposed cracks" do not even exist at this level, whereas the putative remains of the cell boundaries are now visible in even more regions of the specimen, and indeed can now be seen to be a feature of every morphological element of the specimen. The "cracks" that so impressed Bengston and Budd are evidently just surface fractures on the section, of no consequence in any respect. Second, and most important from a scientific point of view, is that a new bilateral feature is revealed at this plane of focus, not visible at the plane shown in Fig. 1G. This is a structure bridging the walls of the coelom on both sides. As we pointed out in (*2*), the fossil section is close to what was

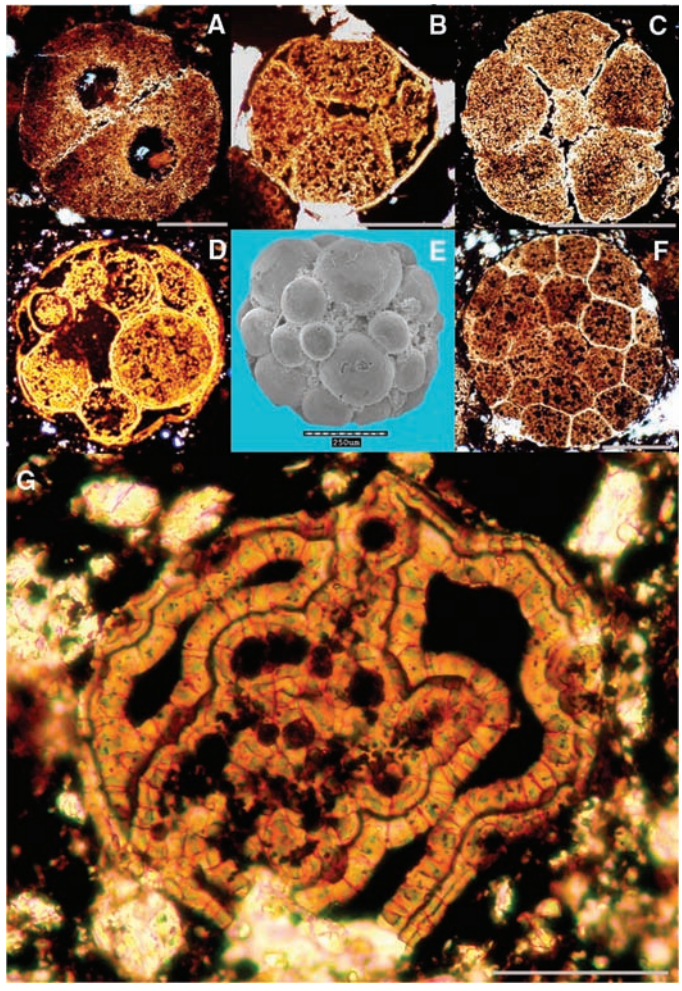


Fig. 1. Images of Doushantuo microfossils showing evidence for cellular preservation. Scale bars represent 200 μm unless otherwise indicated; (D) is similar in scale to (E); for (G), the scale bar represents 40 μm. (**A**) Thin section micrograph of early cleavage embryo showing preservation of nuclear domain. (**B**) Thin section micrograph of early cleavage embryo. (**C**) Thin section micrograph of mid-cleavage embryo. (**D**) Thin section micrograph of mid-cleavage embryo displaying cells of different size. (**E**) SEM image of embryo similar to (D). (**F**) Thin section micrograph of late cleavage embryo. (**G**) Holotype of *Vernanimalcula* under polarized light with crossed nicols.

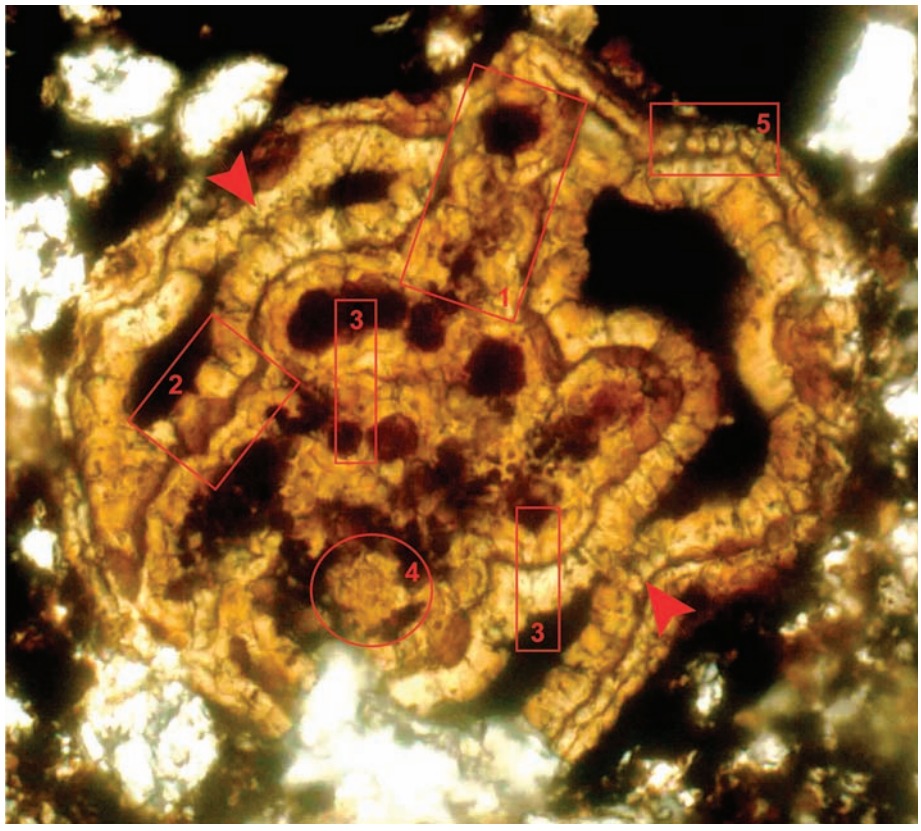


Fig. 2. A deep focal plane image of the holotype specimen of *Vernanimalcula*. Box 1 highlights the pharynx lumen that in this image is clearly continuous between the stomach at the posterior end of the passage and the mouth at the anterior. Box 2 shows a particular view of the juxtaposition of the inner colelomic mesodermal layer and the endoderm. The cuboidal cell-like units of the two layers have clearly different thickness and periodicity. This is typical of biological tissues that perform different functions. The boxes marked 3 are two areas that contain major visible "cracks" in Fig. 1G, but as can be seen here, these "cracks" do not extend to this deeper focal plane. The oval 4 surrounds a clump of globular structures, each of which bears a darker round shadow in the center. The bounded globular structures could be cells, and the darker shadows nuclei. Box 5 highlights a new feature not visible in Fig. 1G: The external layer is also made of repetitive, cell-like units. Note that the dark boundaries between these cell-like units do not invade the inside thick layer surrounding the coelom. The two red arrowheads show a new bilaterally arranged feature not seen before, a structure that crosses the coelom on each side of *Vernanimalcula*. Orientation and dimension are as in Fig. 1.

the ventral surface of the animal; hence, these structures may represent inward bilateral ridges in the floor of the coelomic cavities.

Bengtson and Budd claim that the bilaterally situated pits we noted on the external surface of the fossil are spheroid fans, a common diagenetic form that they illustrate. To make this assertion, however, they assign that identity to indentations on a specimen that we never claimed were surface pits. Perhaps these indentations are spheroid fans; it is irrelevant to the pits in question. Once again, the notion that these features are diagenetic fails to explain the presence within them of distinctively small, regular bounded elements [figure 2A in (*2*)], some of which include the greenish spots discussed above (Fig. 1G).

The main point is that the multiple specimens of *Vernanimalcula* show consistent, bilateral morphology. Bengston and Budd (*1*) imply that we examined 50,000 miscellaneous specimens and picked out these 10 because they had an apparent morphology we were seeking. In fact, as we explained (*2*), the 50,000 microfossils we alluded to were all defined by the presence of recognizable forms, mostly of eggs and embryos. Only the specimens of *Vernanimalcula* have its consistent and particular morphological features. The argument of Bengtson and Budd (*1*) regarding diagenesis is nullified by its own implicit assumption that diagenetic processes would accidentally produce 10 bilaterally symmetrical fossils of similar size and form.

Bengtson and Budd provide images of encrustation within a fossil brachiopod as examples of diagenetic artifacts that are supposed to resemble *Vernanimalcula*. This example is irrelevant, however; it is clearly an error to use deposits on a template of unquestioned biological origin as a model for forms that are supposed to display no biological features. It is impossible to see what they think is similar to *Vernanimalcula* in the morphology of the encrusted brachiopod. Furthermore, it is not obvious that their example even represents what they think it does: The branching structure in the brachiopod could well be the fossilized remains of a fungal organism.

Taphonomy and diagenesis must of course be considered in the analysis of any novel fossil form, but the considerations of Bengtson and Budd provide no answers. They cannot explain the reproducible features, the symmetric morphology, or the internal structural periodicity of the *Vernanimalcula* fossils. Just as it would be a mistake to ignore taphonomy and diagenesis altogether, refusing to look beyond them precludes further exploration and insights into early animal evolution. We confidently predict that many additional specimens of *Vernanimalcula* will be found before long and that they will provide an enhanced view of its anatomy and three-dimensional structure. Discovery of this and other new forms will depend on study of further tens of thousands of specimens.

Jun-Yuan Chen*
Nanjing Institute of Geology and Palaeontology
Nanjing 210008, China
and Laboratory of Pharmaceutical Biotechnology
College of Life Sciences
Nanjing University
Nanjing 210093, China

Paola Oliveri
Eric Davidson
Division of Biology 156-29
California Institute of Technology
Pasadena, CA 91125, USA

David J. Bottjer*
Department of Earth Sciences
University of Southern California
Los Angeles, CA 90089, USA

**To whom correspondence should be addressed.*
E-mail: chenjunyuan@163.net (J.Y.C.); dbottjer@usc.edu (D.J.B.)

References
1. S. Bengtson, G. Budd, *Science* **306**, 1291 (2004); www.sciencemag.org/cgi/content/full/306/5700/1291a.
2. J.-Y. Chen *et al.*, *Science* **305**, 218 (2004); published online 3 June 2004 (10.1126/science.1099213).

2 July 2004; accepted 25 October 2004

ECONOMICS

Markets and Where They Came From

Eric Maskin

There must be dozens of introductory books with the word "microeconomics" in the title, but for ambition alone Samuel Bowles's volume stands out. Not only does Bowles convey the elements of the conventional theory of capitalist economics (albeit in a far less systematic or detailed way than an ordinary textbook), he offers a wealth of cutting-edge material as well. In particular, *Microeconomics: Behavior, Institutions, and Evolution* gives exposure to recent experimental findings that call into question standard behavioral assumptions of economic theory (and generate modifications of these assumptions). It culminates by invoking these modifications along with evolutionary game-theoretic dynamics to explain how the institutions of capitalism came into being.

Microeconomics
Behavior, Institutions, and Evolution
by Samuel Bowles

Princeton University Press, Princeton, NJ, 2004. 598 pp. $49.50, £32.95 ISBN 0-691-09163-3. Roundtable Series in Behavioral Economics.

A standard axiom in economic theory holds that humans are self-interested: if given the choice between helping myself and helping you, I'll favor myself. Economists recognize, of course, that the assumption is not literally true. Many argue, however, that it is good enough for explaining most important economic phenomena. They also cling to it for methodological discipline: were the people in economic models allowed to care about matters other than their own welfare, then almost any behavior could be explained by the easy (but unilluminating) device of giving them a preference for that behavior, or so the fear goes. But Bowles notes that abandoning the self-interest axiom need not lead to complete theoretical permissiveness. The axiom can be replaced by other assumptions that narrow preferences fairly sharply. In any case, he contends that self-interest is a poor approximation of reality.

Bowles puts great weight on experimental results in "public-good" games as evidence for this conclusion. In a two-person version of such games, players are each given, say, $10, which they can then partially invest in a "public project." They choose what portion to invest and keep the rest of the money for themselves. Every dollar contributed to the project results in a $1.50 gross return, split equally between the players. Thus, if both invest their entire $10 in the project, they will each come away with $20 \times \$1.50 \times {}^{1}/_{2} = \$15$. Even so, notice that without some way to cooperate on investments, purely self-interested players should contribute nothing at all—they personally get back only $0.75 for each dollar they contribute (and they don't care about the $0.75 going to the other player). But this prediction is strongly refuted by experiments—subjects typically invest in the project about half the money they are given. Indeed, Bowles maintains that this sort of altruism is an important ingredient in the workings of modern economies.

Civic commerce. Detail from Ambrogio Lorenzetti's fresco *Effects of Good Government in the City* (1338–1340), Palazzo Publico, Siena, Italy.

How else, he asks, but by altruism can we plausibly explain why employees of large companies so often work harder when they own the company themselves? (Each employee is, in effect, participating in a public-good game: working harder to increase the value of the company is personally costly while almost the entire increase accrues to the other employee-owners.)

Yet Bowles goes still further. Not only are altruistic preferences needed for understanding modern economic behavior, they were, he contends, even more important in human prehistory—in particular, for the creation of the institution of private property. As conceived by Bowles, private property is cultural evolution's answer to the problem of wasteful conflict in human production and exchange. In his model, we imagine that there are three types of people (grabbers, sharers, and punishers) and that people are paired up at random to divide a "prize" (a product that the pair jointly produces). A pair of sharers divides the prize equally, as do two punishers, or a punisher-sharer pair. However, a grabber will take the entire prize when paired with a sharer and will fight over it when paired with another grabber. A grabber will also seize the prize from a punisher, who will then enlist fellow punishers to retaliate and wrest the prize back. (These three types correspond to stylized strategies in hunting and foraging interactions before the development of agriculture.)

Bowles shows that, in a dynamic setting (where types with relatively high payoffs, from repeated playing of a divide-the-prize game, proliferate and those with relatively low payoffs diminish in number), there are two limiting configurations toward which the population could evolve. In one—which Bowles calls a Hobbesian equilibrium—the punishers disappear, leaving only grabbers and sharers. In the other, more harmonious, configuration—a Rousseauian equilibrium—the grabbers vanish while punishers and sharers remain. Which configuration emerges depends on the starting point, but Bowles shows the Rousseauian equilibrium is much less stable—not as able to withstand "mutations" (i.e., shocks to the composition of the population)—and therefore less likely to persist over long periods of time.

This sets the stage for property rights. Imagine that each prize is located at a particular site and that there is a fourth type of people, "bourgeois," who behave like grabbers when they control the site (i.e., when they "own" the prize) and like sharers when the other person does. Bowles shows that a band of bourgeois can "invade" a population of grabbers and sharers (because bourgeois types derive higher payoffs from repeated interaction than do the other types) and ultimately drive the others out. Thus, property rights (and bourgeois types) came into existence as a way to avoid fighting and retaliation costs. Of course, such rights rely on the possibility of determining unambiguously who controls a site. This last fact helps explain why they seem not to have emerged before the rise of agriculture; as Bowles observes, it is easier to determine who has possession of cultivated land than of foraging territory.

The reviewer is in the School of Social Science, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540. E-mail: maskin@ias.edu

Published by AAAS

But there is a difficulty with this story: available evidence suggests that the transition to property rights followed a path from Rousseauian (not Hobbesian) equilibrium to bourgeois equilibrium. This is problematic for Bowles's theory because, as noted above, Rousseauian equilibrium and its high level of sharing are relatively fragile. (In particular, they are vulnerable to invasions by sufficiently large bands of grabbers.) One wonders how they survived through the many millennia before bourgeois equilibrium took over. Bowles responds to this difficulty by arguing that altruistic preferences offer a natural way to stabilize Rousseauian equilibrium: if sharing serves not only to avoid conflict but to gratify sharers, the Rousseauian configuration becomes more robust against invasion. Thus, Bowles suggests that altruism may have played a critical role in sustaining Rousseauian behavior in the ages preceding agriculture (and bourgeois types).

To assert that a game-theoretic model with just four strategies can adequately explain the genesis of property rights is bold if not brash, and Bowles acknowledges that his theory is at present only speculative. But, speculative or not, the theory is neat, thought-provoking, and highly original—as is much else in this most unusual take on microeconomics.

EVOLUTION

More Than Metaphorical Power?

Douglas H. Erwin

I began this review while returning from China, one of the world's fastest growing economies. Twelve years ago, when I first visited, China was just beginning to emerge from its economic torpor; today it consumes a majority of the world's cement production and appears to have cornered the world market on construction cranes. Far from a new phenomenon, China's economic strength harkens back to the 1700s when, as Kenneth Pomeranz and other economic historians have shown, China shared many of the advantages of Western Europe. So why did the modern world economy develop in Western Europe instead of China? What factors permit one society to gather the resources sufficient to develop a system that eventually comes to dominate other economies? This conundrum has challenged many historians, who have proposed explanations ranging from geography to environmental endowment and competing polities. In what he calls "an economic analysis of history," Geerat Vermeij, a noted paleontologist at the University of California at Davis, broadens these questions with a novel and intriguing yet at times problematic view of the history of life on Earth.

Nature
An Economic History
by Geerat J. Vermeij
Princeton University Press, Princeton, NJ, 2004. 461 pp. $35, £22.95. ISBN 0-691-11527-3.

Vermeij inquires into the processes that have driven the increased complexity of ecosystems through time as well as the relays of successive dominant groups. His view is that the crucial factor is power, by which he means the acquisition, retention, and use of energy per unit time. Vermeij argues that this same variable is ultimately responsible for success in both natural and human systems. The innovations propelling such successes require the control of sufficient resources to fuel the positive feedback that drives economic expansion. Whether Vermeij's concept of power is sufficient to explain this is the critical question.

Well aware that some will challenge his claim for the generality of an economic approach, Vermeij spends the early part of the book justifying his argument. Because I have long seen economics and economic history as a powerful source of metaphor, I take his point as read, but others may need more convincing. In his view, competition, inequality between units, adaptation, disturbance, trade, and imperfection characterize all economic systems, natural or human, on this planet or any other where life may exist. Vermiej has long advocated the evolutionary importance of escalatory feedbacks between predator and prey, so it comes as no surprise that predation and competition play pivotal roles in his account. Here he again rehearses the strategies of prey and predator, pointing out that the most successful predators exhibit speed, agility, the long-distance detection of prey, and the application of superior force—all reflections of greater power. As a result, competition for resources favors the success of groups with greater and greater power. Through time, such clades wield progressively more influence and replace groups that require fewer resources and less power.

Convincing the reader by the weight of example has a long and honorable tradition—Darwin used it to great effect in *The Origin of Species*. Taking the same approach, Vermiej supports his argument through insight and example by considering predator-prey dynamics, how organisms acquire power (e.g., increases in temperature, metabolic rate, size, and structural complexity), and the roles of nutrients, geography, and other environmental variables. The examples are fascinating, but the lack of quantification is troubling. Statistics were invented, as one wag has it, because the singular of data is anecdote. Selected examples do not allow us to determine whether the case being made is in fact general.

On some topics, the particular focus and lack of a firm quantitative framework lead Vermeij astray. For example, he claims that an oversupply of nutrients is widely held to be the cause of mass extinctions and lesser biotic crises. One wonders "by whom?", for if this was ever widely held it certainly is not today. Anoxia as the extinction trigger has been championed by Tony Hallam and Paul Wignall, but even if anoxia were a general feature of mass extinctions—which is far from clear—nutrient oversupply is not the only cause.

Vermeij is also particularly taken with the role of methane release in mass extinctions, and he discusses this at some length for the end-Permian extinction. Having pioneered this particular hypothesis a decade ago, I am all too aware that the latest, dramatically reduced estimates of the volume of gas hydrates make the case far more problematic than Vermeij would lead one to believe. Although it is easy to quantify estimates for methane release, Vermeij provides no quantitative grounding for his hypotheses. In the end, he argues for a "causal linkage" between massive flood basalts, extraterrestrial impacts, and methane release. Such a meaty stew may be intuitively appealing, but Vermeij provides no clue that most studies indicate that flood basalt eruptions are exceedingly unlikely to be induced by impactors smaller than several hundred kilometers in diameter (versus the 10- to 14-kilometer diameter of the object that triggered the Cretaceous-Tertiary mass extinction) and if any impact occurred at the Permo-Triassic boundary, it was an order of magnitude smaller than required.

Vermeij argues persuasively for the coconstruction of ecology and environment in the creation of power, and in this he is surely right. Others have recently addressed this issue as niche construction or ecosystem engineering, but process-based models remain rudimentary. Greater understanding of many of the patterns that Vermeij documents demands the development of such models and their testing against empirical data from the fossil record. Despite the limitations of Vermeij's approach, *Nature: An Economic History* is not a book easily dismissed. It offers a distinctive point of view and an insightful synthesis that promises to provide the basis of much future work.

The reviewer is at the Department of Paleobiology, National Museum of Natural History, Washington, DC 20560, USA, and the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA. E-mail: erwind@si.edu

INTELLECTUAL PROPERTY

Plants and Intellectual Property: An International Appraisal

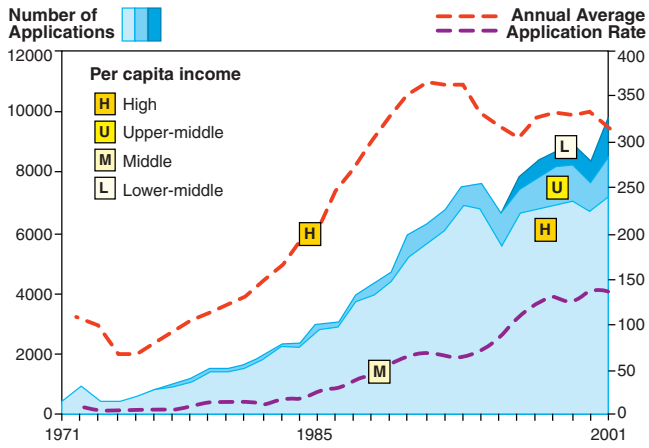
Bonwoo Koo,[1] Carol Nottenburg,[2] Philip G. Pardey[3]*

The era of free and unencumbered access to new crop varieties appears to be passing. This development in intellectual property (IP) has raised a chorus of concerns about the implications for food production and human health, especially throughout the developing world. The downsides of IP have been emphasized by a series of articles in *Science* (*1*–*4*). However, much of the debate occurs in the absence of an understanding of the specifics of the rights available in particular jurisdictions, a practical sense of the rights actually claimed or granted, and their evolution over time. Existing information highlights rich-country developments, with little, if any, attention to developing countries.

Enhanced online at www.sciencemag.org/cgi/content/full/306/5700/1295

While protection of a piece of IP is limited to the countries or regions that grant the protection, international aspects of IP can affect use and especially transfer of the technology or products. The Agreement on Trade-Related Aspects of Intellectual Property (TRIPS), which came into effect in 1995 and is a requirement for members of the World Trade Organization (WTO), inextricably tied trade with patent protection by providing patent owners the right to prevent others from importing a patent product and a product obtained directly from a patented process. Thus, if a producer wants to export a genetically modified crop to a country where there is a patent on the process to make that crop, importation requires the permission of the patent owner.

TRIPS requires that "patents shall be available for any inventions, whether products or processes, in all fields of technology" [Article 27(1)], but also provides added protections for plant varieties by mandating their protection "by patents or by an effective *sui generis* system or by any combination thereof." Sui generis is a term literally meaning "of its own kind" or "unique." Systems for plant variety protection that satisfy the sui generis requirement of TRIPS are often called plant breeders' rights. Although the minimum criteria for patents are set forth in TRIPS, no criteria are elaborated for what constitutes an "effective" sui generis system. There is considerable variation among countries in the implementation and application of these forms of protection. To illustrate the variety of plant-related IP protection on offer worldwide, we describe the different approaches to awarding patents for plants in the United States, Canada, Europe, and the Andean Community; illustrative sui generis systems from the United States (*5*), Europe, and India are contrasted.



Plant breeders' rights. Applications for countries grouped by income (*15*). See notes to Table 1 for criteria used to classify countries. Data for 2002 and 2003 were omitted because of likely underreporting stemming from lags in recording rights claimed or granted.

Utility patents. In most countries, plants and inventions directed to plants or plant products (e.g., seed) are not eligible for a patent. In the United States, however, any living organism that is the product of human intervention (such as by breeding or laboratory-based alteration) is patentable (*6*). In particular, plants and plant parts, including seeds and tissue cultures, have been explicitly held to be patentable (*7*). Plant varieties can also be patented, and, since a recent ruling (*8*), there is no prohibition against obtaining multiple kinds of protection on the same variety. Other plant-related patentable subject matters include plant groups, individual plants and their descendants, plant parts (e.g., specific genes or chromosomes), plant material used in industrial processes, transgenic plants, and particular plant traits.

For patents obtained through the European Patent Office, allowable subject matter is controlled by the European Patent Convention (EPC). Under the EPC, individual plant varieties per se are not patentable; however, claims directed to broader plant groupings are allowable (*9*). Thus, as long as required criteria are met, a claim to "transgenic corn having an insect-resistance gene," for example, is patentable. Plant cells, unlike plant varieties, are patentable because they can result from microbiological processes.

Canadian patent law does not allow patenting of "higher life forms," e.g., plants and animals. In a recent, highly publicized case, *Monsanto* v. *Schmeiser*, the Canadian Supreme Court confirmed this policy but then found that Schmeiser infringed Monsanto's patent claiming a herbicide-resistant gene by growing transgenic canola plants that contained the gene (*10*). Notwithstanding Canadian patent law, this ruling appears to effectively extend Canadian patent protection to plants if they contain a patented gene (*11*).

The Andean Community, a subregional organization made up of Bolivia, Colombia, Ecuador, Peru, and Venezuela, has a common IP regime that is embodied in Decision 486 (*12*), which entered into effect in 2000. Article 20(c) of the Decision expressly prohibits patents on "plants, animals, and essentially biological processes for the production of plants or animals other than non-biological or microbiological processes."

Sui generis systems. It is generally believed that sui generis enables member countries to design their own system of protection for plant varieties as an alternative or addition to a patent system for protecting plants (*13*, *14*). The International Union for the Protection of New Varieties of Plants (UPOV) (*15*)] established a Convention that serves as the basis for sui generis systems

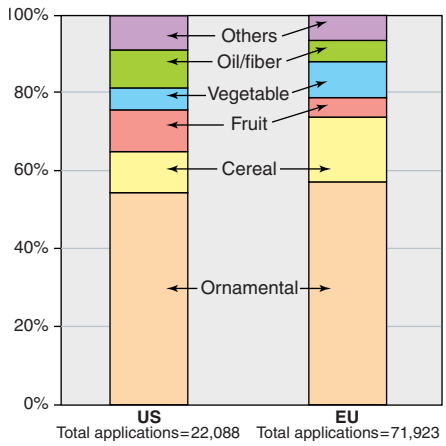
[1]International Food Policy Research Institute, Washington, DC 20006–1002, USA. E-mail: b.koo@cgiar.org. [2]Cougar Patent Law, Seattle, WA 98144, USA. E-mail: c.nottenburg@cougarlaw.com. [3]Department of Applied Economics, University of Minnesota, St. Paul, MN 55108, USA. E-mail: ppardey@apec.umn.edu

*Authorship is alphabetical.

worldwide. Briefly, plant breeders are granted a legal monopoly over the commercialization of their plant varieties (*16*). Notwithstanding this, a number of exemptions from infringement are mandated (use for noncommercial acts, experimental purposes, breeding other varieties) or optional (farmers' saving of seed). Like patents, the rights granted are for a specific time only (not less than 20 years generally or not less than 25 years for trees and vines).

The U.S. Plant Variety Protection Act (PVPA) (*17*) was enacted in 1970, and revised in 1994 to adhere to the 1991 UPOV Convention. The Act provides for protection only for sexually reproduced plants, including first-generation (F_1) hybrids, and tuber-propagated plants (e.g., potato varieties). The counterpart protection for asexually reproduced plants (*18*) is provided by the Plant Patent Act (PPA) enacted in 1930. Probably because it was enacted primarily to benefit the horticulture industry (*19*), the Act protects new and distinct plants that are either invented or discovered, including newly found plant varieties as well as cultivated spores, mutants, hybrids, and newly found seedlings, but excluding tuber-propagated plants. Moreover, implementation of other requirements, such as written description and enablement, for obtaining plant patents is less stringent than for utility patents.

Plant variety protection in the European Union is based on the European Convention (Regulation 2100/94/EC), which in turn is based on the 1991 UPOV Convention. To harmonize and streamline plant variety protection, the Community Plant Variety Right (CPVR) was created in 1995 (*20*). It is not possible to hold simultaneous protection for the same plant variety under both the Community and national system. Furthermore, a CPVR can only be transferred or terminated within all countries of the EU Community and not within selected countries.



Plant breeders' rights stratified by crop categories. United States' data include total number of plant patents granted from 1930 to 2003 and plant variety protection applications from 1970 to 2003. Data for European Union include plant breeders' rights applications to national plant variety offices beginning at or near their inception dates (1942 for Netherlands, 1955 for Germany, 1970s and 1980s for most other countries) to 2003 and applications to the CPVO from 1995 to 2003 (*15*).

As opposed to basing a sui generis system on UPOV, India has chosen a more expansive approach. The Indian Protection of Plant Varieties and Farmers' Rights (PPVFR) Act of 2001 (*21*) ostensibly recognizes the contributions of professional plant breeders and farmers who actively participate in breeding efforts. Thus, the Act contains provisions for "benefit sharing" whereby local communities are acknowledged as contributors of the plants. In a major departure from UPOV, protectable plant varieties include farmers' varieties (those about which there is common knowledge) and other extant varieties including those "notified" under the 1966 Seed Act, or any other variety in the public domain. Furthermore, the Indian PPVFR has some atypical additional requirements for obtaining protection: The applicant must provide information about the origin of the genetic material as well as declare that the variety does not incorporate a restriction technology involving gene(s) that inhibit the development of viable seed.

A provision for farmers' rights in PPVFR [Article 39(iv)] entitles the farmer to save, use, sow, resow, exchange, share, or sell farm produce including seed of a protected variety. Any seeds that are sold, however, cannot be branded. The rights to sell seed appear to undermine the rights of the commercial or farmer breeder. The Act also contains compulsory licensing provisions (*22*), similar to legislation in the United Kingdom (*21*) and in Canada (although such provisions do not pertain to the PVP Act in the United States). Overall, the Indian Act seems to heavily favor "public" over "private" interests. It remains to be seen whether this will qualify as an "effective" sui generis system under TRIPS.

Economies*	Countries (*n*)			
	Statutory protection	Legislation under consideration	Member UPOV	Member WTO
High-income economies (54)	29	4	23	37
OECD (24)	23		21	24
Non-OECD (30)	6	4	2	13
Upper-middle-income economies (37)	20	4	14	28
Lower-middle-income economies (56)	20	18	13	36
Low-income economies (61)	22	3	4	42
Total (208)	91	29	54	143

Plant variety protection legislation worldwide (*30*). Countries are classified into income classes according to World Bank (2004) criteria. High, upper middle, lower middle, and low income are defined as 2003 per capita gross national incomes greater than $9386; $3036–$9385; $766–$3035; and less than $765, respectively. Brackets indicate total number of countries in each income class.

Intellectual Property Landscapes

We have conducted a survey of national IP offices, UPOV, and the WTO (see table at left). Just 91 out of 191 countries surveyed offered statutory IP protection (*23*), while another 29 countries had legislation under consideration. Countries with statutory protection are mostly high- and upper-middle-income countries; less than half the middle- and low-income countries have varietal protection legislation, and most of these are not UPOV member countries.

Although the number of applications by rich countries peaked in the early 1990s, PBR applications filed in upper-middle-income countries have grown steadily since the early 1970s, and the number from lower-middle-income countries only began to rise in the 1980s and is still negligible (*24*) (see figure on first page). From 2000 to 2002, a total of 26,192 PBR applications were lodged worldwide in each country, of which 2909 (11%) were filed in the United States and 11,300 (43%) in European member states of the CPVO, of which nearly one-third were applications made in the Netherlands and more than one-fifth lodged in France (*15*).

The principal proximate cause of the disproportionate activity in developed countries is most likely the lack of rights on offer in poor countries (only 22 of 61 low-income countries have any statutory protection in place for plants). More fundamentally, it probably reflects a range of economic influences regarding the costs and benefits of securing breeders' rights in a particular jurisdiction.

One-third of the PBR applications lodged in 50 UPOV member countries dur-

ing 1998–2002 were lodged by foreigners (see table below). Looking regionally, 31% of the applications in high-income countries were lodged by foreigners, 65% in upper-middle-income countries, 25% in lower-middle-income countries, and 38% in low-income countries. The variation is even more apparent in individual countries; for example, the share of applications filed by foreigners is 85% in Switzerland, 42% in the United States, 24% in Japan, and 11% in France (*25*). This substantial fraction of foreign applications indicates extensive potential spillovers of varietal improvement research done in one locale on seed market and production developments elsewhere in the world.

	Applications (*n*)		
Economies	**Total**	**Residents**	**Nonresidents**
High-income economies (23)	39,079	26,893	12,186
Upper-middle-income economies (11)	5,583	1,945	3,638
Lower-middle-income economies (12)	6,109	4,592	1,517
Low-income economies (4)	487	299	188
Total (50)	51,258	33,729	17,529

Share of plant breeder rights applications lodged by foreigners, 1998–2002. See table 1 for country income classification criteria. Bracketed figures indicate number of countries included in the data. [Source (*33*)]

The percentage of plant patents and PBRs granted to different plant groups is similar in the United States and Europe (see figure opposite, top). Ornamental crops account for more than half the total applications in both the United States and Europe (*15*), while cereal crops (such as wheat and corn) is the next biggest group (11% in the U.S. and 17% in Europe). Other major groups of plants that are protected are oil and fiber plants, fruit crops, and vegetables. Because plant-related utility patents are a comparatively recent phenomenon in the United States, only 5% of all plant-related protection are utility patents, of which 55% pertain to corn and 40% to soybeans (*15*).

Conclusions

International treaties like TRIPS and intergovernmental organizations like UPOV leave scope for much variation in the specifics of plant IP protection. Our review of national plant variety legislation shows that countries are exploiting these degrees of freedom, presumably tailoring plant IP legislation to local circumstances. Variations include such fundamentals as the types of IP offered, species and genera encompassed, costs, and extent of farmers' rights.

The long-term effects of these variations on the rate and direction of plant innovation are yet to be determined. Although the geographical scope of protection is expanding, IP markets are still quite segmented—the preponderance of protection pertains to rich-country jurisdictions, leaving poor countries free to tap these technologies. Moreover, a sizable share of the protected varieties are ornamentals, not food crops, and most plant varieties are afforded protection that enables rights holders to limit or exclude others from marketing but not breeding with the protected material. In addition, the lion's share of food staples produced in developing countries are consumed where grown and are not exported to rich countries (*26*). Thus, concerns that IP rights are currently limiting the freedom to research or commercialize developing-country food staples seem overstated. Misplaced concerns over IP seem to be diverting policy attention from more fundamental negative trends, notably, the slowdown in investment in agricultural R&D worldwide, especially the research targeted to poor people's food crops, and deteriorating domestic capacities—during the past decade in particular—to conduct agricultural R&D in many poor countries, especially throughout sub-Saharan Africa (*27*).

None of this is to deny that possible increases in the transaction costs of moving plant material from one IP jurisdiction to another may be slowing international spillovers, but the IP effects per se are more likely to reduce technological spillovers from poor to rich countries, rather than germ plasm flows in the other direction. Moreover, any slowdowns may be temporary. Harmonizing plant IP legislation is likely to lower these transaction costs (one variant of this being the formation of Europe's Community Plant Variety Office); increased knowledge of the details of national legislation is another avenue for improving efficiencies in the international movement of plant innovations subject to intellectual protection. In addition, the disclosure and information requirements coupled with increasing Internet access may help streamline and progress breeding efforts. At the very least, more complete examination and investigation of these changing IP landscapes internationally should be undertaken before bold assertions about the consequences of IP are taken as truths.

References and Notes

1. M. Heller, R. Eisenberg, *Science* **280**, 698 (1998).
2. D. Kennedy, *Science* **302**, 357 (2003).
3. R. C. Atkinson *et al.*, *Science* **301**, 174 (2003).
4. R. Beachy, *Science* **299**, 473 (2003).
5. The United States has awards called utility patents for any kind of plant and plant patents—more akin to plant breeder rights—for certain asexually reproduced plants.
6. *Diamond* v. *Chakrabarty* 447 US 303, 1980.
7. *Ex parte Hibberd* 227 USPQ 433 (Bd. Pat. App. & Int. 1985).
8. *J.E.M. AG Supply* v. *Pioneer Hi-Bred International* 122 S. Ct. 593 (2001).
9. Directive 98/44/EC (effective in all European Union member states 30 July 2000) also aims to harmonize protection for biotechnological inventions (including plant protection) among the European Union members (*28*).
10. *Monsanto Canada Inc.* v. *Schmeiser*, 2004 SCC 34.
11. C. Nottenburg, "*Schmeiser* v. *Monsanto*" in *Navigating the Patent Maze*, posted 22 July 2004, www.cougarlaw.com.
12. Decision 486, Common Intellectual Property Regime, Andean Community; available at www.comunidadandina.org/ingles/treaties/dec/D486e.htm.
13. http://www.upov.int/en/about/upov_convention.htm.
14. The Convention was established in 1961 and has been revised three times; not all UPOV member countries are bound by the latest Convention.
15. Further information can be found in the supplemental material.
16. Protection confers the right to exclude others from producing or reproducing, propagating, offering for sale, selling or other marketing, exporting, importing or stocking for any of the above purposes the protected variety [Article 14(1) of the Convention].
17. www.ams.usda.gov/science/PVPO/PVPO_Act/PVPA.htm.
18. Plant Patent Act, 35 U.S.C. §§161–164 (1930).
19. C. Fowler, P. Mooney. *Shattering: Food, Politics, and the Loss of Genetic Diversity* (Univ. of Arizona Press, Tuscon, 1990).
20. Community Plant Variety Office (CPVO). *Annual Report, 2002* (Community Plant Variety Office, Paris, 2003); available at www.cpvo.eu.int/default.php?res=1&w=820&h=543&lang=en&page=droit/legislation.htm.
21. P. Brahmi *et al.*, *Curr. Sci.* **86**, 392 (2004).
22. Statutory Instrument 2002 No. 247 The Patents and Plant Variety Rights (Compulsory Licensing) Regulations, 2002.
23. Only 54 out of a total of 191 (28%) of the countries surveyed having legislation are members of the UPOV Convention.
24. Over time, some countries with PBRs conforming to the 1978 UPOV Convention have relaxed restrictions on the scope of crop protection offered. In China, for instance, a total of 10 species were eligible for protection in September 1999, growing to 30 species by March 2002 (including 5 major cereals, 2 oil crops, 2 roots and tubers, 10 vegetables and fruits, and 11 flowers and grasses but excluding cotton) (*29*).
25. P. G. Pardey, B. Koo, C. Nottenburg, "Creating, protecting, and using crop biotechnologies worldwide in an era of intellectual property," *Minn. J. Law Sci. Technol.* (in press).
26. E. Binenbaum *et al.*, *Econ. Dev. Cultur. Change* **51**, 309 (2003).
27. P. G. Pardey, N. M. Beintema, *Slow Magic: Agricultural R&D a Century After Mendel* (IFPRI Food Policy Report, International Food Policy Research Institute, Washington, DC, 2001).
28. europa.eu.int/eur-lex/pri/en/oj/dat/1998/l_213/l_21319980730en00130021.pdf.
29. B. Koo *et al.*, An Option Perspective on Generating and Maintaining Plant Variety Rights in China. Department of Applied Economics Staff Paper P03-8 (Univ. of Minnesota, St. Paul, December 2003).
30. Data were compiled from on-line searches of national IP offices, (*31*), and (*32*).
31. www.upov.org/en/about/members/index.htm.
32. www.wto.org.
33. UPOV, "Plant variety protection statistics for the period 1998–2002" (C/37/7, International Union for the Protection of New Varieties of Plants, Geneva, 2003).

34. The authors thank the International Food Policy Research Institute, CAMBIA, and the University of Minnesota for financial support; and D. Ashton, J. Sharples, E. Castelo-Magalhães, and H. Wright for their help in preparing this paper.

PLANETARY SCIENCE

Alien Weather at the Poles of Mars

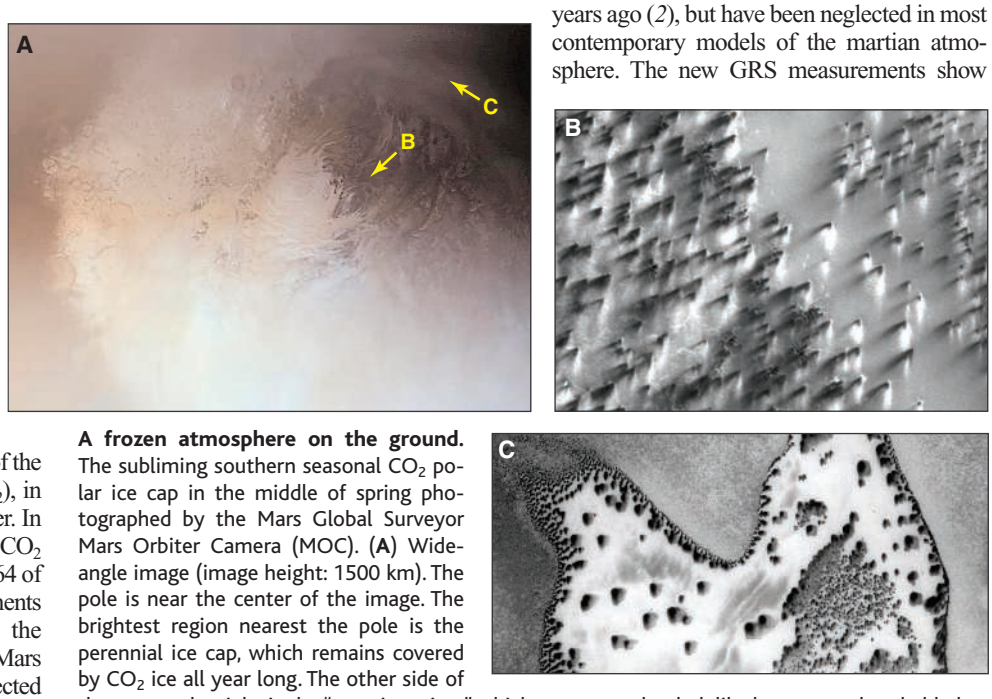
François Forget

In many respects, the weather on Mars is very similar to the weather on Earth. On both planets, the Hadley circulation (the process that generates the trade winds) is important at low latitudes, whereas "baroclinic" planetary waves (a succession of low- and high-pressure zones) dominate the weather system at mid-latitudes. Of course, Mars is colder and dryer than Earth, and water clouds are less important on Mars than they are on Earth. Conversely, martian mineral dust lifted by the winds is not easily scavenged from the atmosphere and tends to strongly affect the opacity of the thin atmosphere and its thermal structure. Despite these differences, the martian weather usually corresponds to what one would predict for a planet that is a colder, drier, desertlike Earth. However, there is one aspect of martian meteorology that has no terrestrial counterpart. This is the direct condensation of the main constituent of the martian atmosphere, carbon dioxide (CO_2), in the polar regions during autumn and winter. In a step toward understanding the martian CO_2 cycle, Sprague *et al.* (*1*) report on page 1364 of this issue their analysis of direct measurements of the martian atmosphere taken by the Gamma Ray Spectrometer (GRS) aboard Mars Odyssey. Their analysis reveals unexpected fluctuations in atmospheric composition that create weather with no equivalent on Earth.

As much as 30% of the martian atmosphere condenses every year to form polar caps in both hemispheres, inducing large surface-pressure variations over the entire planet. At first glance, this phenomenon may seem straightforward, but numerous observations by the NASA Mars Global Surveyor and Mars Odyssey missions, and most recently by the European Space Agency Mars Express spacecraft, suggest that this event is very complex. As is often the case in Mars exploration, the more we observe this phenomenon, the more puzzling it becomes. Many aspects of traditional meteorology and of cloud and snow microphysics must be reinvented to understand the Mars CO_2 cycle. The analysis by Sprague and colleagues reveals a new facet of the CO_2 cycle on Mars: an increase in atmospheric argon over the southern polar regions in autumn followed by its dissipation during winter and spring.

At first glance, the Sprague *et al.* findings are expected. CO_2 condenses on the martian surface to form CO_2 ice, whereas argon and the other noncondensable gases—principally nitrogen (N_2) and oxygen (O_2)—comprising 5% of the martian atmosphere do not. The magnitude of these events, however, is usually underestimated. Indeed, Sprague *et al.* (*1*) now show that the mean argon mixing ratio is enhanced by as much as a factor of 6 during winter and depleted by a factor of 2 to 3 during spring. As a result, the air composition varies strongly with location and season. Indeed, noncondensable gases constitute up to 30% of the bulk southern polar atmosphere during the winter solstice (and probably much more locally) compared to about 5% on average over the planet. Under such conditions, the partial pressure of CO_2 would be lower than expected, the CO_2 frost point temperature would be decreased by several kelvin, and the surface thermal infrared cooling would be reduced by more than 5%. More importantly, because the mean molecular weight of noncondensable gases is only 32.3 g mol^{-1} (as compared with 44 g mol^{-1} for CO_2), the enrichment of such gases near the surface where most of the CO_2 condenses would induce deep static instability and vertical mixing.

These aspects were considered by Hess 25 years ago (*2*), but have been neglected in most contemporary models of the martian atmosphere. The new GRS measurements show that the winter martian atmosphere is characterized by a sizable latitudinal gradient of different molecular weight gases that form a deep layer at the edge of the polar vortex. Meteorologists have not previously had to consider such density gradients, although a close analogue would be the gradient of salinity in oceans that oceanographers have had to incorporate into their calculations. In practice, the enrichment of lighter noncondensable gases observed during the winter solstice would have the same affect on the atmospheric circulation as a 13 K temperature gradient (as used in the traditional thermal wind equation, for instance). This gradient would tend to reduce the intensity of the polar vortex, and to



A frozen atmosphere on the ground. The subliming southern seasonal CO_2 polar ice cap in the middle of spring photographed by the Mars Global Surveyor Mars Orbiter Camera (MOC). (**A**) Wide-angle image (image height: 1500 km). The pole is near the center of the image. The brightest region nearest the pole is the perennial ice cap, which remains covered by CO_2 ice all year long. The other side of the cap on the right is the "cryptic region," which appears to be dark like bare ground probably because the cap consists of transparent slab ice (*8*). During the spring in these regions, the ice tends to sublime from the bottom, forming vents that allow pressurized gases to escape. (**B**) This process erodes the surface along "spider-shaped" (*9*) ragged channels that can be seen through the ice, and the ejected material falls downwind to form fans. (**C**) In some other locations, the ejected material forms curious dark spots of puzzling aspect (*7*). (B) and (C) are MOC narrow-angle images (image height: 2.95 km). The global heterogeneity of the subliming cap may be related to its formation during the polar night. The three pictures were obtained a few days apart.

CREDIT: NASA/JPL/MALIN SPACE SCIENCE SYSTEMS

The author is at the Laboratoire de Meteorologie Dynamique, Institut Pierre Simon Laplace, Universite Paris 6, Paris cedex 5, France. E-mail: francois.forget@lmd.jussieu.fr

Published by AAAS

favor the transport of noncondensable gases outside the polar region.

Another aspect of the martian polar night atmosphere that is far from understood is the formation of CO_2 ice clouds and snowfall. Although it is thought that most of the carbonic ice directly condenses on the surface, a fraction should also condense in the atmosphere, strongly influencing the radiative properties of the atmosphere and the martian surface (*3*). Most of these clouds form in the polar night, and thus evidence of their existence has remained theoretical (*4*) or indirect (*3*). It is only with the advent of the Mars Global Surveyor laser altimeter MOLA—which acts as a light detection and ranging (LIDAR) instrument—that a variety of cloud shapes varying over space and time have been observed (*5*). There have been several attempts to model the complex behavior of these clouds, which seem to form in topography-induced updrafts, buoyancy waves in the lee of mountains, or even in exotic convection cells (*4*, *6*, *7*). One difficulty is that, because CO_2 is the major constituent of the atmosphere, the microphysics of martian CO_2 ice clouds is unlike that of any clouds on Earth or on other planets of the Solar System.

At the end of the polar night, condensation stops, but the behavior of CO_2 ice does not become simple (*8*). The sublimation of the frozen atmospheric layer is characterized by spectacular albedo changes (*9*) and explosive gas eruptions that erode the surface year after year (*10*), forming curious dark spots of multiple shapes (*8*) (see the figure). In the southern hemisphere each year, a large part of the cap (the so-called cryptic regions) remains quite transparent and dark and rapidly sublimates (see the figure). In contrast, other areas at the same longitude become very bright and ultimately outlast the summer to form the perennial CO_2 ice cap at the south pole (*9*). This geographical distribution still has not been explained. Furthermore, the existence of the perennial CO_2 ice cap, a relatively thin (*11*) frozen atmospheric reservoir near the south pole, is puzzling. Any changes in its albedo or the evolution of the planet's orbital parameters (which are highly variable) would make the CO_2 ice cap either disappear or grow much bigger within a few years. Somewhere hidden in the alien meteorology that controls the formation of the martian CO_2 ice cap, there must be some stabilizing feedbacks that remain to be discovered.

References

1. A. L. Sprague *et al.*, *Science* **306**, 1364 (2004); published online 7 October 2004 (10.1126/science.1098496).
2. S. Hess, *J. Geophys. Res.* **84**, 2969 (1979).
3. F. Forget *et al.*, *J. Geophys. Res.* **100**, 21219 (1995).
4. F. Forget *et al.*, *Icarus* **131**, 302 (1998).
5. G. H. Pettengill, P. G. Ford, *Geophys. Res. Lett.* **27**, 609 (2000).
6. G. Tobie *et al.*, *Icarus* **164**, 33 (2003).
7. A. Colaprete *et al.*, *J. Geophys. Res.* **108**, 5081 (2003).
8. H. H. Kieffer, *Sixth International Conference on Mars*, Pasadena, California, 20 to 25 July 2003 (LPI, Houston, TX, 2003), p. 3158.
9. H. H. Kieffer *et al.*, *J. Geophys. Res.* **105**, 9653 (2000).
10. S. Piqueux *et al. J. Geophys. Res.* **108**, 3-1 (2003).
11. J. P. Bibring *et al.*, *Nature* **428**, 627 (2004).

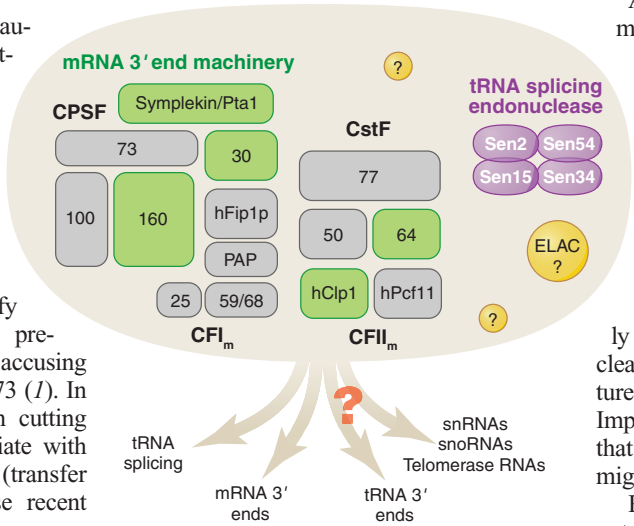
MOLECULAR BIOLOGY

Knives, Accomplices, and RNA

Marvin Wickens and Tania N. Gonzalez

The weapon is missing and the authorities are frustrated. The attacks are simple—a phosphodiester bond is severed cleanly to form the 3′ end of messenger RNA (mRNA), which is then free to receive its tail of polyadenosine [poly(A)]. The attacking gang of proteins is large and well known, but an intensive search has been mounted to identify the culprit that actually cuts the pre-mRNA. A recent study points an accusing finger at one gang member, CPSF73 (*1*). In another clue, proteins involved in cutting pre-mRNAs also physically associate with proteins that cleave pre-tRNAs (transfer RNAs) during splicing (*2*). These recent findings suggest surprising links among proteins that cut different types of RNA.

Primary RNA transcripts are cleaved by endonucleases to generate the 3′ ends of mRNAs, tRNAs, microRNAs, and certain small nuclear and nucleolar RNAs (snRNAs snoRNAs). The formation of mRNA 3′ ends can be reconstituted in the test tube (*3*–*6*), and in mammalian cells at least 14 different proteins are required for this process (see the first figure). The factors CPSF and CstF recognize the critical sequences in pre-mRNAs, whereas other factors are required for the cleavage step and for addition of the poly(A) tail. All of these factors, except the poly(A) polymerase (PAP), are complexes containing multiple proteins. Although 900 kD worth of factors have been isolated, it is unclear whether the enzyme that actually cuts pre-mRNA is among them.



A Swiss Army endoknife? The tRNA splicing endonuclease (purple) and mRNA 3′-end machinery (green and gray) associate with one another. Proteins in purple associate with those in green (*2*); proteins in gray have not been tested. The complex may process other types of RNAs, and contain other endoribonucleases (yellow). [mRNA factors adapted from a figure by W. Keller (Biozentrum, Basel)]

A serendipitous clue to the identity of the mRNA endonuclease comes from studies of tRNA processing. Mutations in the *ELAC2* gene appear to cause susceptibility to prostate cancer (*7*). ELAC2 is an endonuclease that cleaves 3′ extensions from pre-tRNAs (*8*) in mammalian cells, like its close relatives in plants and Archaea (*9*). The ELAC proteins are similar in sequence to the 73-kD subunit of CPSF (CPSF73) (*7*, *10*). CPSF73 is a member of a subfamily of metallo-β-lactamase enzymes that cleave nucleic acids using a distinctive structure that coordinates two zinc ions (*11*). Implicit in these reports (*8*–*11*) is the idea that if ELACs cut pre-tRNAs, then CPSF73 might cut pre-mRNAs.

Ryan *et al.* (*1*) recently showed that the putative active site of CPSF73 is essential for viability of yeast cells. Moreover, mRNA 3′ cleavage in vitro, long thought to be metal independent, is stimulated by zinc, consistent with CPSF73 being the perpetrator (*1*). Yet the case against CPSF73 is open: It is unclear whether mRNA 3′ cleavage is defective in CPSF73 mutants or whether CPSF73 is even a nuclease. Complicating matters further, the *Drosophila* zinc-finger endoribonuclease, Clipper, is related to a different CPSF subunit (*12*, *13*).

The notion that CPSF73 is the enzyme that forms mRNA 3′ ends is seductive, in part because the two steps that form both

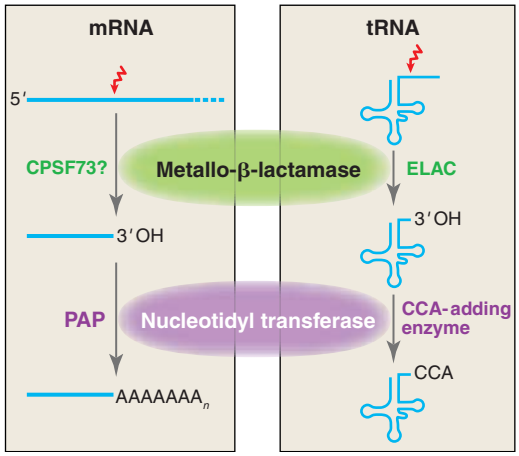
The authors are in the Department of Biochemistry, University of Wisconsin, Madison, WI 53706, USA. E-mail: wickens@biochem.wisc.edu

mRNA and tRNA 3′ termini would then be strikingly similar (see the second figure). Pre-mRNAs and pre-tRNAs first are cleaved by related metallo-β-lactamase enzymes, leaving a 3′-hydroxyl group. Then, two related nucleotidyl transferases—CCA-adding enzyme for tRNA, and PAP for mRNA—add nontemplated RNA sequences [CCA and poly(A)] to the new ends (*10*). The presence of tRNA-like structures at the ends of certain viral genomes may be an evolutionary relic of this process (*14*).

In a recent study, Paushkin *et al.* (*2*) reveal a startling collusion among proteins involved in cleaving mRNA and tRNA precursors. The tRNA splicing endonuclease of yeast consists of four subunits: Sen2, Sen34, Sen54, and Sen15 (see the first figure) (*15*). Sen2 cleaves the 5′ splice site, and Sen34 the 3′ splice site of tRNAs. Using sequence comparisons, Paushkin *et al.* (*2*) identified candidate tRNA splicing endonucleases in human cells, then isolated the splicing complexes with tagged proteins. As expected, complexes containing all four subunits were detected and accurately cleaved pre-tRNAs. Unexpectedly, these tRNA splicing complexes also contained factors involved in the formation of mRNA 3′ ends. Among these mRNA 3′ cleavage factors were HsClp1, CstF64, and symplekin. These proteins are required for 3′ end formation in vitro (*16*, *17*), and associate with one another in cell extracts (*18*). Remarkably, yeast mutants with a mutation in the symplekin homolog, PTA1, are defective in tRNA splicing (*19*), providing an independent link between tRNA splicing and mRNA processing.

The association between the tRNA splicing endonuclease and the mRNA 3′-end machinery may be dynamic. Most components of the mRNA machinery may be able to interact with the tRNA endonuclease (see the first figure), although they may not all reside in a single complex. Indeed, the proportion of different mRNA 3′ end factors associated with the tRNA endonuclease varies considerably. It will be critical to determine whether the isolated tRNA endonuclease accurately cleaves pre-mRNAs.

Might the enzyme that cuts tRNAs also produce the 3′ ends of mRNAs? The "forensic" evidence says no. Whereas mRNA 3′ end cleavage leaves a 3′-hydroxyl and 5′-phosphate (*20*, *21*), tRNA splicing endonucleases leave a 5′-hydroxyl and 2′,3′-cyclic phosphate (*15*). If tRNA-specific splicing endonucleases cut mRNA at all, they probably do not produce normal polyadenylated mRNAs. Most likely, the tRNA and mRNA endonucleases carry out separate duties even though they are found together. In human cells where tRNA splicing endonuclease activity has been reduced by RNA interference, uncleaved pre-mRNAs accumulate (*2*). This defect in mRNA 3′ end processing implies a functional link between mRNA and tRNA maturation. This defect may reflect a collapse of the endonuclease complex, loss of mRNA cleavage activity, or indirect effects.



Parallels in mRNA and tRNA 3′ end processing. Two pairs of related enzymes may act consecutively to form the 3′ ends of mRNAs and tRNAs.

Regulated mRNA splicing of Sen2, the tRNA splicing endonuclease subunit, may coordinate the abundance of certain tRNAs and mRNAs. Human Sen2 pre-mRNA is alternatively spliced such that exon 8 is either included (Sen2) or omitted (Sen2Δ8). The Sen2Δ8 protein appears to associate less efficiently with other tRNA endonuclease subunits, and to differ in its association with mRNA 3′ end factors (*2*). Moreover, complexes containing Sen2Δ8 conceivably may act on non-tRNA substrates because they cut pre-tRNAs with altered specificity.

Perhaps multiple endoribonucleases assemble into a Swiss Army endoknife—a complex of distinct blades, each honed to attack different RNAs. The mRNA 3′-end machinery alone may cut many other types of RNA. For example, formation of the 3′ ends of certain snRNAs, snoRNAs, and perhaps yeast telomerase RNA requires certain proteins that also are needed to cut pre-mRNAs (*22*, *23*). ELAC, the enzyme that processes tRNA 3′ ends, might well be in the complex, partnered with the tRNA splicing endonuclease. And what of Drosha, the endonuclease that processes the 3′ ends of microRNAs? Or the activity that forms the ends of nonadenylated histone pre-mRNAs? This hypothetical endonuclease complex may be analogous to the exosome, a complex of 3′-to-5′ exoribonucleases (*24*).

Currently, the only activities known to be associated with one another are the tRNA enodonuclease and several mRNA 3′ end processing factors. Might these interactions be artifactual? Only a small fraction of the mRNA cleavage factors appear to be associated with the tRNA splicing enzymes, and in vitro it appears that tRNA processing enzymes are not required to cleave mRNA 3′ termini. Yet, a direct assessment of tRNA splicing proteins in mRNA cleavage factor preparations seems warranted. Termination of transcription from both mRNA and snoRNA genes requires common proteins, suggesting another link in the biogenesis of different types of RNA (*25*). Moreover, translation, transcription, and mRNA processing factors all form higher order assemblies that are dispensable for minimal reactions in vitro. RNA cleavage factors may have similar complexities to reveal.

The observation that endoribonucleases associate with one another, even transiently, prompts new views of the cellular assaults on RNAs, and a new search for the weapons cache. If diverse endoribonucleases roam the cell together, then sorting out which blade cuts which RNA will require much care, including site-directed interrogation of each suspect. CPSF73 has not yet been convicted of pre-mRNA 3′ cleavage, but is taking the stand even as tRNA splicing endonucleases are revealed as accomplices. These proteins, along with other shadowy members of endoribonuclease complexes, may yet be tried for crimes against other types of RNAs.

References

1. K. Ryan *et al.*, *RNA* **10**, 565 (2004).
2. S. V. Paushkin *et al.*, *Cell* **117**, 311 (2004).
3. D. F. Colgan, J. L. Manley, *Genes Dev.* **11**, 2755 (1997).
4. M. Edmonds, *Prog. Nucleic Acid Res. Mol. Biol.* **71**, 285 (2002).
5. L. Minvielle-Sebastia, W. Keller, *Curr. Opin. Cell Biol.* **11**, 352 (1999).
6. J. Zhao et *al.*, *Microbiol. Mol. Biol. Rev.* **63**, 405 (1999).
7. S. V. Tavtigian *et al.*, *Nature Genet.* **27**, 172 (2001).
8. H. Takaku *et al.*, *Nucleic Acids Res.* **31**, 2272 (2003).
9. S. Schiffer *et al.*, *EMBO J.* **21**, 2769 (2002).
10. L. Aravind, *In Silico Biol.* **1**, 69 (1999).
11. I. Callebaut *et al.*, *Nucleic Acids Res.* **30**, 3592 (2002).
12. C. Bai, P. P. Tolias, *Mol. Cell. Biol.* **16**, 6661 (1996).
13. S. M. Barabino *et al.*, *Genes Dev.* **11**, 1703 (1997).
14. N. Maizels *et al.*, Eds., *The RNA World* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1999), pp. 79–112.
15. J. Abelson *et al.*, *J. Biol. Chem.* **273**, 12685 (1998).
16. H. de Vries *et al.*, *EMBO J.* **19**, 5895 (2000).
17. J. Zhao *et al.*, *Mol. Cell. Biol.* **19**, 7733 (1999).
18. Y. Takagaki, J. L. Manley, *Mol. Cell. Biol.* **20**, 1515 (2000).
19. J. P. O'Connor, C. L. Peebles, *Mol. Cell. Biol.* **12**, 3843 (1992).
20. C. L. Moore *et al.*, *EMBO J.* **5**, 1929 (1986).
21. M. D. Sheets *et al.*, *Mol. Cell. Biol.* **7**, 1518 (1987).
22. S. Dheur *et al.*, *EMBO J.* **22**, 2831 (2003).
23. M. Morlando *et al.*, *Mol. Cell. Biol.* **22**, 1379 (2002).
24. P. Mitchell, D. Tollervey, *Nature Struct. Biol.* **7**, 843 (2000).
25. E. J. Steinmetz, D. A. Brow, *Mol. Cell. Biol.* **23**, 6339 (2003).

MATERIALS SCIENCE

Shaping Crystals with Biomolecules

James J. De Yoreo and Patricia M. Dove

The shapes of crystals found in biomineral structures such as the skeletons of marine organisms differ dramatically from those of crystals grown in pure solution. The formation of these complex, often hierarchical structures is difficult to reconcile with the simple mechanistic model of crystal growth by step propagation across crystallographic faces (often referred to as the terrace-ledge-kink model) (*1*). Two decades ago, researchers developed the stereochemical recognition model (*2*, *3*), which holds that these shapes are stabilized through the binding of peptides and proteins to otherwise unstable faces, presumably because the stereochemical match to the crystal lattice lowers their surface energies. Recent studies show how the two models can be reconciled.

The stereochemical recognition paradigm found support in a series of investigations of the macroscopic shapes of calcium carbonate crystals. Some of the crystals were grown in the presence of proteins extracted from carbonate biominerals; others were grown with organic additives that mimic those proteins. The extracted proteins were highly acidic, often containing large fractions of aspartic acid residues (*3*). The resemblance of the carboxylate group of aspartic acid to the carbonate group of calcite and the strong affinity of carboxylates for calcium provided a rational basis for the conjectured stereochemical match. In addition, the presence of proteins in the mineral phases suggested that they actively control crystal growth. In several cases, specific geometric relationships between the modifier and a particular crystal face were proposed that suggested an excellent stereochemical match. Even detailed molecular modeling studies focused on such arrangements (*4*).

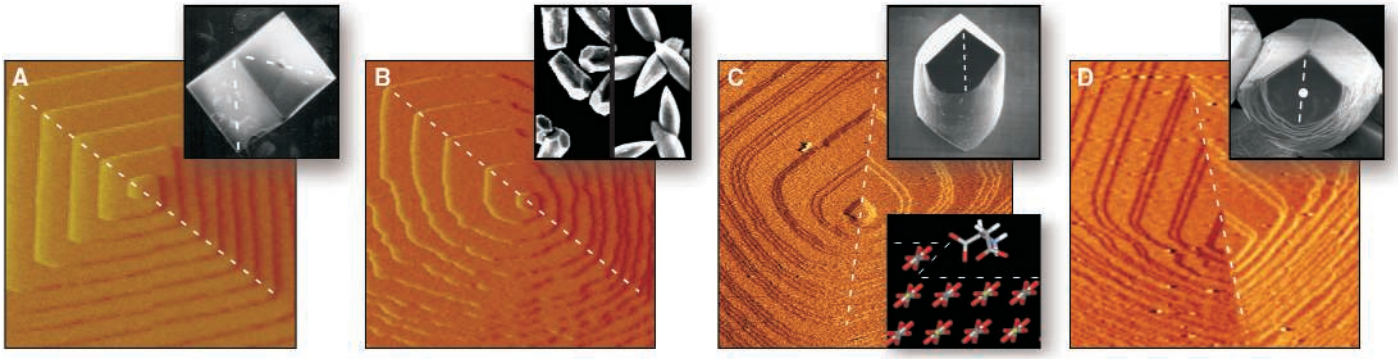
However, the paradigm depended heavily on macroscopic examination of crystal shapes. The subtle links between atomic-scale dynamics and macroscopic expression were not examined. Moreover, the term "stabilization" suggested a reduction in surface free energy—yet the newly expressed faces were typically rough and nonplanar, indicating that they were not stable in a thermodynamic sense. Finally, even in pure systems, crystal shapes are as much an expression of growth kinetics as they are of equilibrium energetics. And growth kinetics is controlled at kinks on atomic ledges, not on flat faces (*1*).

Over the past decade, a series of studies has examined the atomic-scale dynamics of crystallization. They explored the growth of crystal surfaces in a variety of systems including carbonates, oxalates, phosphates, and phthalates, with the use of growth modifiers composed of small inorganic molecules (*5*–*8*), moderate-sized organic molecules (*9*, *10*), amino acids (*11*), peptides (*10*), and proteins (*9*, *12*–*14*) (see the figures). These studies have led to a new understanding of shape modification.

We first look at the effect of magnesium, a major constituent of seawater that plays a key role in carbonate modification (*15*). Magnesium preferentially inhibits step motion at the corners of growth hillocks (see the first figure, panels A and B). This effect should flatten the corners of the calcite rhombus, leading to elongation and roughening of the crystal shape (*5*). Recent nucleation and growth experiments have confirmed this shape evolution (*16*) (first figure, panel B).

Calcite shape can also be modified by acidic amino acids. When individual enantiomers are added to calcite growth solutions, the glide-plane symmetry of calcite is broken. The resulting crystal shape reflects both the chirality of the amino acid and the step-specificity of the interaction (first figure, panel C) (*11*). The shape change of the resulting macroscopic crystal mimics that of the elementary steps (first figure, panel C). Molecular models show that step edges provide the most favorable binding environment and give binding energies that exhibit the same asymmetry seen both at the atomic scale and in the macroscopic crystal (*11*).

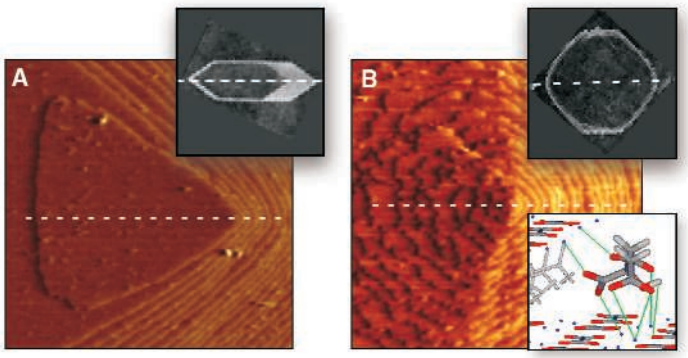
A more complex system is represented by the growth of calcium oxalate monohydrate, a pathological biomineral, in the presence of citrate, a naturally occurring inhibitor and therapeutic agent. Citrate is a nonplanar molecule with three carboxylate groups and a hydroxyl group. Its impact on step morphology or kinetics depends strongly on the type of face (*9*) (see the second figure). Molecular modeling shows that several factors—including calcium spacing, step riser angle, hydroxyl-to-oxalate distance, and electrostatic interactions—determine the magnitude of the binding energy.



How crystals change shape I. Studies of calcite show that changes in crystal shape are directly related to modification of step morphology for a wide range of modifiers. Each panel shows an atomic force microscope image of growth hillocks (main image) and a scanning electron microscope image of the crystal shape (inset), with dashed lines to indicate the glide plane symmetry. (**A**) Pure calcite. (**B**) Calcite plus Mg^{2+} (*5*); inset shows crystals grown at Mg^{2+}:Ca^{2+} mole ratios of 1.5 (left) and 2.0 (right) [after (*16*)]. (**C**) also shows D-aspartic acid binding to a partiular step on calcite (*11*). (**D**) Calcite plus AP8 protein extracted from abalone nacre [after (*14*)]. (C) also shows D-aspartic acid binding to a particular step on calcite (*18*).

J. J. De Yoreo is in the Chemistry and Materials Science Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA. E-mail: deyoreo1@llnl.gov P. M. Dove is in the Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. E-mail: dove@vt.edu

Published by AAAS



How crystals change shape II. Similar changes as in the first figure can be seen for calcium oxalate monohydrate crystals. (**A**) Pure calcium oxalate monohydrate. (**B**) Calcium oxalate monohydrate plus citrate (*9*). Insets after (*21*). (B) also shows minimum energy configurations for citrate binding to a particular step from molecular simulations (*9*).

The change in macroscopic crystal shape again mimics the change in growth hillock shape (second figure, panel B).

Finally, closer to the in vivo situation, we consider how calcite crystal shapes are modified by the introduction of AP8, a protein extracted from abalone nacre (*14*). Even for this complex modifier, the changes in atomic-scale morphology are step-specific and directly determine the shape of macroscopic crystals (first figure, panel D) to give morphologies that are quite similar to those caused by simple amino acids and polypeptides (*11, 12, 14*).

Although the mechanisms of growth modification are diverse, the source of shape change in these studies is clear: Crystal shape is controlled by step-specific interactions between growth modifiers and individual step edges on preexisting crystal faces. The common appearance of new, rough, rounded surfaces, which clearly are not faces, is not a result of stereochemical matching to a particular atomic plane of the crystal. Rather, changes in the elementary step shape generate a similarly modified bulk crystal shape through the self-replicating process of crystal growth.

The emergence of new faces is thus a macroscopic manifestation of the kinetics caused by molecular-scale interactions at the step edges. In this way, the terrace-ledge-kink model (*1*) merges smoothly with the concept of stereochemical recognition proposed two decades ago.

Subsequent to the development of the stereochemical recognition model, a growing body of evidence has shown that the shape of biominerals is often controlled through molding of solid or gelated amorphous precursors (*17–19*). Nonetheless, many biomineral structures present clear evidence for active control during crystal growth. Furthermore, the concept can be used to aid the design and synthesis of crystalline materials.

Despite this new level of understanding, one mystery remains. How are changes affecting elementary steps on one face translated into the emergence of adjacent faces? The answer seems to lie at the corners between faces. Here, steps from adjacent faces must converge in regions of high curvature. This convergence provides an opportunity for surface energy terms associated with curvature to become important and for steps to bunch, either for energetic reasons or as a result of kinetic fluctuations (*20*). The behavior of steps in this regime remains to be explored. Such studies should provide the final piece to the puzzle of shape modification.

References

1. W. K. Burton, N. Cabrera, F. C. Frank, *Philos. Trans. R. Soc. London* **243**, 299 (1951).
2. S. Mann *et al.*, *Science* **261**, 1286 (1993).
3. L. Addadi, S. Weiner, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4110 (1985).
4. A. L. Rohl, *Curr. Opin. Solid State Mater. Sci.* **7**, 21 (2003).
5. K. J. Davis, P. M. Dove, L. E. Wasylenki, J. J. De Yoreo, *Am. Mineral.* **89**, 714 (2004).
6. U. Becker, B. Gasharova, *Phys. Chem. Miner.* **28**, 545 (2001).
7. L. A. Touryan, M. J. Lochhead, B. J. Marquardt, V. Vogel, *Nature Mater.* **3**, 239 (2004).
8. T. N. Thomas, T. A. Land, J. J. De Yoreo, W. H. Casey, *Langmuir* **20**, 7643 (2004).
9. S. R. Qiu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1811 (2004).
10. S. Guo, M. D. Ward, J. A. Wesson, *Langmuir* **18**, 4284 (2002).
11. C. A. Orme *et al.*, *Nature* **411**, 775 (2001).
12. D. A. Walters *et al.*, *Biophys. J.* **72**, 1425 (1997).
13. J. B. Thompson *et al.*, *Biophys. J.* **79**, 3307 (2000).
14. G. Fu, S. Valiyaveettil, B. Wopenka, D. E. Morse, in preparation.
15. H. Elderfield, G. Ganssen, *Nature* **405**, 442 (2000).
16. Y. J. Han, J. Aizenberg, *J. Am. Chem. Soc.* **125**, 4032 (2003).
17. E. Beniash, J. Aizenberg, L. Addadi, S. Weiner, *Proc. R. Soc. London Ser. B* **264**, 461 (1997).
18. S. Raz, S. Weiner, L. Addadi, *Adv. Mater.* **12**, 38 (2000).
19. L. B. Gower, D. J. Odom, *J. Cryst. Growth* **210**, 719 (2000).
20. E. D. Williams, N. C. Bartelt, *Science* **251**, 393 (1991).
21. Y. Shirane, S. Kagawa, *J. Urol.* **150**, 1980 (1993).

PLANETARY SCIENCE

How Neptune Pushed the Boundaries of Our Solar System

Alessandro Morbidelli

Planetary scientists are finding increasing evidence that the orbital separations between the giant planets increased substantially as a result of interactions between the planets and a disk of "planetesimals" that were left over after planet formation. The evidence comes from the Kuiper belt, a population of small (diameter <1000 km) bodies at the outer edge of today's Solar System that are the last remnants of this disk.

The author is in the Observatoire De la Cote d'Azur, B.P. 4229, 06034 Nice Cedex 4, France. E-mail: morby@obs-nice.fr

Twenty years ago, computer simulations (*1*) showed that planetary orbits should expand or contract to conserve energy and angular momentum while the planets eject the planetesimals left over from planet formation from their neighborhoods. A decade later, Malhotra (*2*) proposed that the orbits of Pluto and the "Plutinos"—a subpopulation of the Kuiper belt—were probably caused by Neptune's migration. She concluded that Neptune, which is now at 30 astronomical units (AU; 1 AU is the mean distance from the Earth to the Sun), has moved outward by at least 7 AU since its formation.

The orbital period of the Plutinos is 1.5 times that of Neptune, a behavior referred to as a 2:3 resonance. As the orbit of Neptune expanded, the orbital period of the planet increased. Hence, the location of the 2:3 resonance with Neptune also moved outward through the planetesimal disk (see the figure). Malhotra (*2*) showed that when planetesimals were swept by the resonance like house dust by a broom, they were likely to be "trapped" in resonance. Trapped planetesimals then moved outward with the resonance, while the ellipticity of their orbits slowly increased. In contrast, untrapped planetesimals kept their original orbital radius and small orbital ellipticity and inclination. According to this picture, the current Plutinos are the trapped planetesimals.

However, the study was not definitive. The proportions of Plutinos with high and low orbital inclinations could not be reproduced in Malhotra's migration model (*3*). Furthermore, important properties of the Kuiper belt, such as its lower-than-expect-

ed mass (*4*), and the two types ("hot" and "cold") of Kuiper belt objects other than the Plutinos (*5*) could not be explained solely by the migration model. They seemed to suggest that some other mechanism sculpted the Kuiper belt.

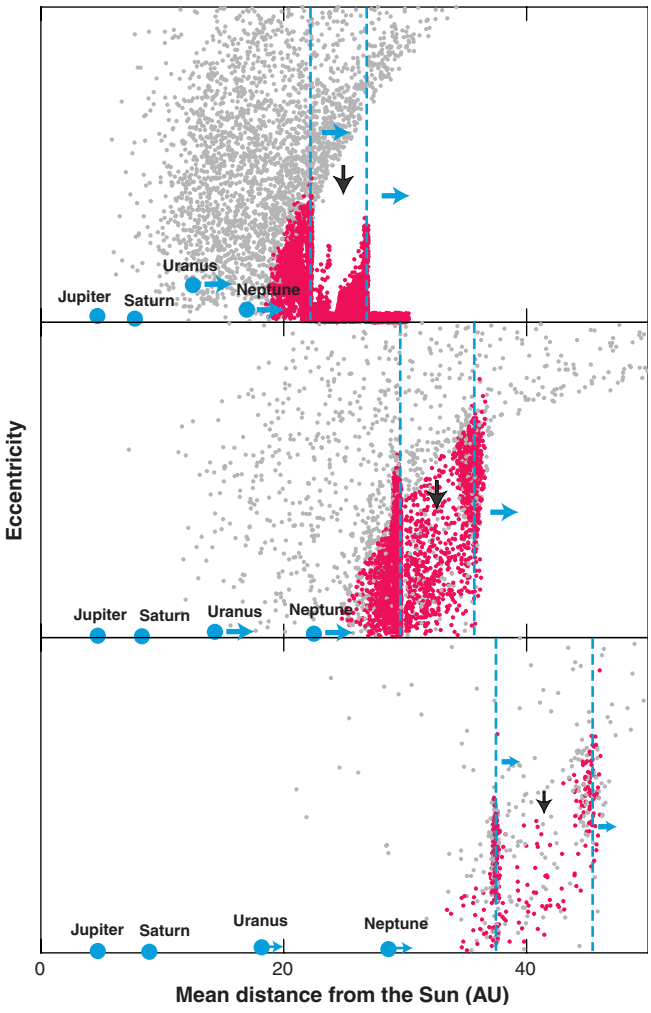
A series of more sophisticated models (*6*–*8*) of the migration process now provide confirmation of the original idea, showing that all the main properties of the Kuiper belt can be explained by planet migration alone. But for this to be possible, the planetesimal disk through which the migration took place must not extend beyond about 30 AU.

The assumption of an outer bound for the original planetesimal disk may appear to be arbitrary, but it is not. The Kuiper belt, which is the last remnant of the disk, ends at about 50 AU, where the period of the objects is exactly double that of Neptune (1:2 resonance) (*9*). The planetesimal disk was thus truncated. But where was its original outer edge? None of the mechanisms proposed to explain the truncation of the planetesimal disk attributes any role to the planets (*10*). It is therefore intriguing that the Kuiper belt ends at the location of a resonance with Neptune. This observation suggests that the outer edge of the planetesimal disk was originally well inside 50 AU, and that Neptune's migration pushed the Kuiper belt beyond the disk's original boundary (see the figure).

There are two reasons to believe that the primordial location of the outer edge of the disk was close to 30 AU. First, it explains why Neptune's migration stopped there. In our model (*7*), the planet tends to reach the outer edge of the disk and stop there. A significantly more extended disk would have driven Neptune beyond its current position (*7*). Second, it explains the current small mass of the Kuiper belt. If the region of space now inhabited by the belt was originally empty, the current mass of the belt reflects the fraction of the disk planetesimals that were pushed there during Neptune's migration. This fraction was presumably small, because most planetesimals were ultimately ejected from the Solar System.

To date, two mechanisms have been identified to push a small fraction (about 0.1%) of the disk's planetesimals beyond the original disk edge and implant them on stable Kuiper belt orbits. The first explains the "hot" Kuiper belt objects and a part of the plutino population, while the second explains the "cold" Kuiper belt objects.

According to the first mechanism, Neptune scattered the planetesimals with which it had close encounters as it moved through the disk. Some planetesimals suffered multiple encounters and were transported outward on elliptic, inclined orbits (gray dots in the figure). A small fraction of these objects still exists today, forming what is usually called the "scattered disk." Occasionally, some scattered disk objects entered a resonance with Neptune. Resonances can modify the ellipticity of the orbits. If the ellipticity is decreased, the sequence of encounters stops and the body becomes "decoupled" from Neptune, like a Kuiper belt object. Due to Neptune's migration, some of the decoupled bodies escaped from the resonances and became permanently trapped in the Kuiper belt (*8*). These bodies preserved the large inclinations acquired during the Neptune-encountering phase and now form the "hot" Kuiper belt objects. A few scattered objects also reached stable Plutino orbits (*8*). When the latter are put together with the Plutinos generated by Malhotra's mechanism, one obtains a very good match with the observed orbital distribution of the Plutinos, thus removing the problems discussed in (*3*).



The orbital evolution of the outer Solar System. The three panels show sketches of the beginning, middle, and end of planetary migration, based on model results in (*7*–*9*). The vertical axis denotes the eccentricity (a measure of the ellipticity of the orbits). The vertical dashed lines show the locations of the 2:3 (left) and 1:2 (right) resonances with Neptune. The blue arrows indicate the direction of migration of Uranus, Neptune, and the resonances. The disk planetesimals are colored, depending on whether they have had close encounters with Neptune (gray) or not (red). Gray objects should have a wide range of orbital inclinations, whereas red objects should preserve their original small inclination. Most gray objects form the "scattered disk," but a few decrease their orbital ellipticity and mix with the red objects (black arrow). The planetesimal disk is originally truncated at about 30 AU (**top panel**). As Neptune moves outward, some objects are transported beyond this boundary. At the end (**bottom panel**), a small fraction of surviving objects are left in the Kuiper belt, which is approximately bounded by the 2:3 and 1:2 resonances with Neptune.

According to the second mechanism, while Neptune was migrating through the disk, its 1:2 and 2:3 resonances swept through the disk, capturing a fraction of the disk planetesimals [as in Malhotra's theory (*2*)]. When the 1:2 resonance passed beyond the edge of the disk, it continued to carry its load of objects. Because Neptune's migration was presumably not a perfectly smooth process, the resonance gradually dropped objects during its outward motion, all along its way up to its final position at about 50 AU (*6*) (see the figure). This process explains the current location of the outer edge of the Kuiper belt. Because the 1:2 resonance does not enhance the orbital inclinations by much, the bodies transported by the resonance preserved their initial small inclination, forming the "cold" Kuiper belt objects.

Most properties of the Kuiper belt can thus be explained by invoking planet migration in a truncated planetesimal disk. Does this mean that the true evolution of the outer Solar System has finally been uncovered? Probably not. One only has to look at the Moon to realize that something important is missing in all of the scientists' scenarios. The dark spots on the Moon are huge impact basins that formed some 700 million years after the

Published by AAAS

Moon itself, during a cataclysm usually called the "late heavy bombardment" (*11*). What caused this bombardment, which occurred throughout the Solar System?

Previous studies (*12*) argued that a massive planetesimal disk in the outer Solar System could have caused the late heavy bombardment. In my opinion, there are not many realistic alternatives to this explanation. However, according to our current understanding, after the migration of the giant planets was over, the Solar System looked essentially like the current one, with no massive planetesimal populations left. It is thus tempting to conjecture that the late heavy bombardment was triggered by a late start of planet migration. But why did migration start late, rather than soon after planetary formation? The answer is not known, yet.

References and Notes

1. J. Fernandez, W. Ip, *Planet. Space Sci.* **44**, 431 (1984).
2. R. Malhotra, *Astron. J.* **110**, 420 (1995).
3. R. S. Gomes, *Astron. J.* **120**, 2695 (2000).
4. The current mass of the Kuiper belt is only 0.01 to 0.1 Earth masses, but a few tens of Earth masses had to exist in the region where the Kuiper belt bodies formed in order for the growth time scale to be short enough. Therefore, the current Kuiper belt apparently contains only 0.1 to 1.0% of its primordial mass.
5. The inclination distribution of the nonresonant Kuiper belt objects is bimodal (*13*). A first component, called "cold population," has inclinations smaller than about 4°. A second component, called "hot population," has a much broader inclination distribution, extending up to 30° to 40°. The hot and cold populations seem to have different physical properties (*14, 15*).
6. H. F. Levison, A. Morbidelli, *Nature* **426**, 419 (2003).
7. R. S. Gomes, A. Morbidelli, H. F. Levison, *Icarus* **170**, 492 (2004).
8. R. S. Gomes, *Icarus* **161**, 404 (2003).
9. C. Trujillo, M. Brown, *Astrophys. J.* **554**, 95 (2001).
10. Four plausible explanations have been proposed: (i) The outer part of the disk was destroyed by the passage of a star (*16*). (ii) It was photo-evaporated, owing to the presence of massive stars in the neighborhood of the Sun (*17*). (iii) The gas disk was extended, but planetesimals beyond some threshold distance could not grow because of the enhanced turbulence in the outer disk (*18*). (iv) Distant dust particles and/or planetesimals migrated to a smaller heliocentric distance during their growth as a consequence of gas drag (*19, 20*).
11. W. K. Hartmann, G. Ryder, L. Dones, D. Grinspoon, in *Origin of the Earth and Moon*, R. Canup, K. Righter, Eds. (Univ. of Arizona Press, Tucson, AZ, 2000), pp. 493–512.
12. H. F. Levison, L. Dones, C. Chapman, A. Stern, M. Duncan, K. Zahnle, *Icarus* **151**, 286 (2001).
13. M. Brown, *Astron. J.* **121**, 2804 (2001).
14. H. F. Levison, S. A. Stern, *Astron. J.* **121**, 1730 (2001).
15. C. A. Trujillo, M. E. Brown, *Astrophys. J.* **566**, 125 (2002).
16. S. J. Ida, J. Larwood, A. Burkert, *Astrophys. J.* **528**, 351 (2000).
17. F. C. Adams, D. Hollenbach, D. Laughlin, U. Gurti, *Astrophys. J.* **611**, 360 (2004).
18. J. M. Stone, C. F. Gammie, S. A. Balbus, J. F. Hawley, in *Protostars and Planets IV*, V. Mannings, A. P. Boss, S. S. Russell, Eds. (Univ. of Arizona Press, Tucson, AZ, 1998), pp. 589–612.
19. A. N. Youdin, F. H. Shu, *Astrophys. J.* **580**, 494 (2002).
20. S. Weidenschilling, in *Comets II*, M. Festou, H. U. Keller, H. Weaver, Eds. (Univ. of Arizona Press, Tucson, AZ, in press).

STRUCTURAL BIOLOGY

Voltage Sensor Meets Lipid Membrane

Roderick MacKinnon

Electrical impulses propagate rapidly along the membranes of living cells. The molecular components that make this possible are proteins known as voltage-dependent ion channels. These channels open in response to changes in the voltage across the cell membrane, and it is precisely this voltage-dependent property that allows them to propagate electrical impulses. Thirty years ago, Armstrong and Bezanilla demonstrated that when voltage-dependent ion channels experience a change in the membrane voltage, tiny electrical charges known as gating charges move relative to the membrane electric field (*1*). This fundamental observation suggested that a transmembrane voltage change exerts an electric force on the gating charges, causing the pore within the channel protein to open.

Now we know that voltage-dependent potassium ion (K^+) channels contain an ion-selective pore domain with a gate, and a voltage sensor domain (segments S1 to S4) attached to the pore (see the figure). The gating charges correspond to positively charged arginine residues located on the otherwise hydrophobic S4 segment. Thus, voltage-sensing results from a repositioning of the arginine residues within the membrane electric field that is associated with structural rearrangements of the voltage sensor, and these structural rearrangements are linked to the opening of the pore's gate (*2, 3*).

The voltage sensor's structure and the process by which the gating charges are repositioned have been subjects of intense controversy. On the basis of electrophysiological studies, a number of structural models of the voltage-dependent K^+ channel have been proposed. These models share the feature of an S4 helix that is isolated from the lipid membrane by a protein wall consisting of helices S1, S2, and S3 on the channel's lipid-facing perimeter (*4–8*). They posit that a voltage change across the membrane causes a translation or rotation of the S4 helix, which would move the S4 helix and its positively-charged arginine residues within an aqueous "gating pore." Recently, x-ray crystal structures (*9*), biotin-avidin accessibility studies (*10*), and electron microscopy (*11*) of KvAP, a prokaryotic voltage-dependent K^+ channel, have suggested a different model. In this model (the paddle model) the voltage sensor is a highly mobile domain, and it is "inside-out" in the sense that helices S1, S2, and S3 do not isolate S4 from the membrane; instead, S4 itself is located at the protein-lipid interface. Specifically, S4 engages part of S3 to form a helix-turn-helix "paddle" that could somehow move at the protein-lipid interface. It is the location of S4 that is at the center of the controversy. Is S4 at the protein-lipid interface or is it shielded from the lipid membrane by S1, S2, and S3? The paddle model is based on a collection of data—a full-length crystal structure with obvious distortions of its voltage sensor (*12*), a crystal structure of the isolated voltage sensor (*13*), and accessibility data (*10*)—and thus is conceptual, not atomic, and in many respects still needs to be defined.

A recent report in *Science* by Perozo and his co-workers (*14*) presents new data on the structure of the KvAP voltage sensor. These authors studied the spin-label side-chain accessibility and mobility of KvAP K^+ channels in lipid membranes using electron paramagnetic resonance (EPR) spectroscopy. This is a particularly informative technique for analyzing membrane proteins because it uses accessibility parameters determined from the spectral effects of lipid-soluble (O_2) and water-soluble (NiEDDA) relaxing agents to distinguish between lipid-accessible and water-accessible surfaces (*15*). A spin-label side chain at a specific position on a protein can thus be classified into one of three categories: buried beneath the protein surface, on the surface exposed to aqueous solution, or on the surface exposed to lipid. Furthermore, a side-chain mobility value provides additional information; surface positions tend to have a higher mobility value than those buried inside the protein.

All voltage-dependent ion channels undergo conformational changes. Which conformation did Perozo and his colleagues analyze? The KvAP voltage sensor is held in a closed conformation when the voltage is negative (for example, –100 mV) on the inner membrane surface relative to the out-

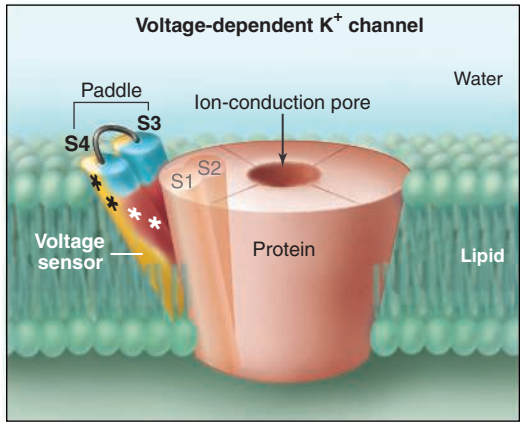
The author is in the Howard Hughes Medical Institute and Laboratory of Molecular Neurobiology and Biophysics, Rockefeller University, New York, NY 10021, USA. E-mail: mackinn@rockefeller.edu

side surface of the cell. The sensor moves to its open conformation upon membrane "depolarization" to 0 mV, causing the pore to open. After a few seconds at 0 mV, the pore becomes inactivated and ion conduction stops by an as yet unknown mechanism, although in all likelihood the voltage sensor remains roughly in its open conformation. This is the condition under which the EPR experiments were carried out by Perozo's group: membranes at 0 mV and the voltage sensor presumably open.

What did they find? In a spin-label scan of the voltage sensor (helical segments S1 to S4), the pattern of accessibility satisfies expectations for a protein that spans the membrane several times, that is, the water-exposed residues occur in the hydrophilic loops between transmembrane segments. This result is presented as being inconsistent with the paddle model, and the inconsistency is demonstrated through accessibility calculations made on an atomic version of a paddle model. This analysis is inappropriate because the paddle model, as originally presented, contained uncertainties (particularly with respect to the positions of S1 and S2) that precluded calculations of atomic coordinates (*10*).

Nevertheless, the new data convey a very telling message: Side-chain mobility and lipid accessibility both increase as measurements proceed from S1 to S4. As the authors point out, this suggests that S1 and S2 are buried in a protein environment, whereas S3 and S4 are more exposed to the lipid membrane (see the figure).

Side-chain accessibility and mobility data in the absence of distance constraints are insufficient to deduce a complex protein structure. But the authors propose a structural interpretation based on the assumption that the crystal structure of the isolated voltage sensor (*13*) approximates a native conformation in the membrane lipid bilayer. Given this assumption, they then ask which orientation of the crystal structure of the isolated sensor and pore best satisfy the constraints imposed by the spin-label data? Perozo and colleagues conclude that S1 and S2 must represent the surface that lies against the pore, away from the membrane, thus accounting for their buried location. They conclude that S3 and S4 helices reside on the outer perimeter of the protein against the membrane, thus accounting for their high lipid exposure and high mobility. In reality, the voltage sensor in situ may adopt a somewhat different structure than that of the isolated voltage sensor. This is especially true at its covalent attachment site (S4 of the voltage sensor to S5 of the pore) where differences are implied by the spin-label data, and at the surface formed by S1 and S2, which somehow forms an interface with the pore. But the basic conclusion seems inescapable: S3 and S4 reside at the protein-lipid interface, against the lipid membrane (see the figure).



A model of the KvAP K^+ channel with a voltage-sensor at the protein-lipid interface. Accessibility and mobility data mapped onto a representation of the K^+ channel show the relation of the channel's voltage sensor to the lipid membrane, water surfaces, and pore (*14*). The figure depicts one possible explanation for the accessibility data reported by Perozo and colleagues (*14*), with two exposed and two buried positive charges in an apparently open conformation of the voltage sensor. Four identical subunits (brown) of the voltage-dependent K^+ channel surround a central ion-conduction pore. Each subunit has a voltage sensor comprising α-helical segments S1 to S4. Accessibility measurements suggest that S1 and S2 are closest to the pore. In contrast, S3 and S4 reside at the protein-lipid interface, with one surface (yellow) exposed to lipid, another (red) exposed to protein, and with water (blue) coating the "top." The S4 helix contains four positively charged arginine residues (asterisks) that allow the channel to sense changes in membrane voltage.

In the crystal structures of KvAP, the carboxyl-terminal half of S3 and S4 form the helical-hairpin paddle of the voltage sensor (*9*). Although such an analysis was not presented in the paper, when the spin-label data are mapped onto the paddle a striking pattern constrains the orientation of the paddle in the membrane. As shown in the figure, one face is exposed entirely to lipid (yellow), whereas the other is exposed to water near the hairpin turn (blue) and to a mixture of protein and lipid (red and yellow) further "down," as though the paddle's long axis were roughly perpendicular to the membrane plane with one face toward the channel and the other toward the membrane.

Where are the arginine residues? The authors state that four out of six are buried. Yet the answer to this question with respect to the positive charges involved in gating requires careful consideration. In the Shaker voltage-dependent K^+ channel isolated from fruit fly, the amino-terminal four S4 arginine residues contribute to the gating charge, whereas the arginine residues near the carboxyl-terminal end of S4 do not (*16, 17*). The crystal structures of KvAP provide an explanation for this: The gating-sensitive arginines correspond precisely to those located on the voltage-sensor paddle, whereas the gating-silent arginines are found off the paddle, on the so-called S4 to S5 linker (*9*). Of the four paddle arginines, the spin-label data suggest that two (the third and fourth) are buried within the protein and two (the first and second) are exposed on the surface. The first of these is exposed only to lipid, whereas the second is exposed to lipid and water (see the asterisks in the figure).

The new data from Perozo and his colleagues are entirely consistent with the concepts proposed for the open conformation of the paddle model (*10*). Many structural aspects of the voltage-gated K^+ channel are still uncertain. For instance, the new spin-label data by Perozo's group do not address the question of how the voltage sensor moves. Fluorescence measurements of the *Shaker* K^+ channel have led to the hypothesis of small (~2 Å) charge movements across a strong electric field that is highly focused by aqueous crevasses penetrating the protein surface (*3*). In contrast, avidin accessibility experiments on the KvAP channel suggest large (at least 15 Å) movements of the voltage-sensor paddle at the protein-lipid interface (*10*). There will no doubt be further disagreements over movements of the voltage sensor, and these will drive our understanding still further. What we need next are new structures, additional biochemical and functional analyses, accessibility data on the closed conformation of the channel, and a better chemical understanding of the protein-lipid interface. All of these are sure to come our way in the near future.

References

1. C. M. Armstrong, F. Bezanilla, *J. Gen. Physiol.* **63**, 533 (1974).
2. F. J. Sigworth, *Q. Rev. Biophys.* **27**, 1 (1994).
3. F. Bezanilla, E. Perozo, *Adv. Protein Chem.* **63**, 211 (2003).
4. R. Horn, *J. Gen. Physiol.* **120**, 449 (2002).
5. C. S. Gandhi, E. Y. Isacoff, *J. Gen. Physiol.* **120**, 455 (2002).
6. F. Bezanilla, *J. Gen. Physiol.* **120**, 465 (2002).
7. H. P. Larsson, *J. Gen. Physiol.* **120**, 475 (2002).
8. M. Laine *et al. Neuron* **39**, 467 (2003).
9. Y. Jiang *et al., Nature* **423**, 33 (2003).
10. Y. Jiang *et al., Nature* **423**, 42 (2003).
11. Q. Jiang *et al., Nature* **430**, 806 (2004).
12. Protein Data Bank (PDB) accession code 1ORQ.
13. PDB accession code 1ORS.
14. L. G. Cuello *et al., Science* **306**, 491 (2004); published online 15 October 2004 (10.1126/science.1101373).
15. W. L. Hubbel *et al., Curr. Opin. Struct. Biol.* **8**, 649 (1998).
16. S. A. Seoh *et al., Neuron* **16**, 1159 (1996).
17. S. K. Aggarwal, R. MacKinnon, *Neuron* **16**, 1169 (1996).

INTRODUCTION

Higher Standards

This special issue of *Science* looks at the development of precision measurement, how its tools have been developed and adapted for better performance, and how the standards used today may be further improved. Historically, measurements were often based on somewhat arbitrary local units. In his Viewpoint, Ashworth (p. 1314) describes, from a British perspective, the development of a standardized metrology as applied to weights and measures and how the burgeoning commerce of the industrial revolution drove its development.

Arguably, today's most important commodity is time. Atomic clocks keep time with an accuracy of about 1 part in 10^{15} and already have shown applications in everyday life, from navigation to satellite and high-speed optical communications. With a view to improving on this, Diddams *et al*. (p. 1318) review the progress made with time and frequency standards in the optical regime, where the higher operating frequency offers the possibility of a finer time scale. [See also the Report by Margolis *et al*. (p. 1355), who show that the transition frequency of a strontium ion can now be determined to within 1 hertz, and the Research Article by Marian (published on *Science* Express) on using optical frequency combs for precision spectroscopy.]

Exploiting the recognition that the ideal way to count quantities would be in fundamental discrete units, or quanta, was severely hampered by the lack of technology. Standards of the fundamental units were all based on experimental artifacts. With improved technology, combined with improved materials and ingenious engineering, Flowers (p. 1324) reviews the recent progress toward quantum-based standards. Quantum mechanics is one of the most sophisticated tools available. Giovannetti *et al*. (p. 1330) review theoretical and experimental work on beating the so-called quantum limit. By playing tricks with quantum mechanics, such as using quantum squeezing, they show how measurement can be improved on.

CREDIT: K. BUCKHEIT/*SCIENCE*

The News pages present a spectrum of the cutting edge of metrology. Andrew Watson describes efforts to turn measurement standards into a counting game: counting atoms to define a kilogram, electrons to define current and capacitance, phonons for temperature, and photons for luminosity (p. 1308). Watson also describes the difficulties facing the semiconductor industry in measuring and reproducing features on chips now that they are down to a few nanometers in size (p. 1309). Atomic clocks are already incredibly accurate, but Robert F. Service found a group that is pushing the decimal point further back with some help from quantum entanglement (p. 1310). The European satellite Hipparcos is the premier metrologist of the heavens, plotting the positions of stars to unprecedented accuracy. Its plot of the position of the Pleiades, however, proved to be a headache, as Govert Schilling found out (p. 1312). In an effort to get a glimpse of lightning-fast processes inside the atom, researchers are striving to create shorter and shorter laser pulses to provide a flash gun. The fastest pulses are now measured in attoseconds: a billionth of a billionth of a second. But, asks Alexander Hellemans (p. 1313), how do you measure a pulse that short?

– IAN OSBORNE AND DANIEL CLERY

CONTENTS

See also Report on page 1355 and Science Express Research Article by Marian.

Science

Published by AAAS

NEWS

Measurement and the Single Particle

It's a guiding principle in science: to understand something, break it down to its smallest constituent parts. Now metrologists are getting into that game by counting out individual atoms, electrons, photons, or phonons to put their standards on rock-solid foundations

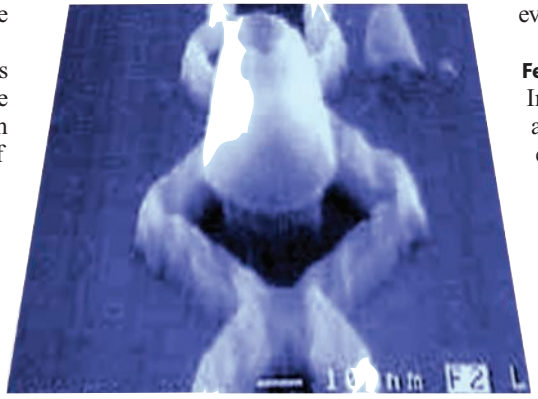
In a vault in Paris sits an object that haunts metrologists' dreams. A cylinder of gleaming platinum alloy about the size of a mobile phone, it tells the world exactly how much mass makes up a kilogram—the only fundamental standard still defined by a physical object. Measurement experts consider the reference kilogram an annoying anachronism. Most would love to scrap it in favor of, say, a definition based on a set number of gold atoms. First, though, they must learn to count atoms reliably—a project still in the works (*Science*, 7 May, p. 812).

Similar efforts are under way in labs around the world, as technicians of the extremely small count everything from electrons to photons and phonons in hopes of bringing measurements down to basics. "It's turning all measurements into a counting exercise," says Patrick Josephs-Franks of Britain's National Physical Laboratory (NPL) in Teddington, near London.

Consider the electron, the lifeblood of electronic circuits. Neil Zimmerman of the U.S. National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland, Mark Keller of NIST in Boulder, Colorado, and colleagues have developed an electron pump that dispenses single electrons into a capacitor and, once the capacitor's voltage change has been measured, counts them back out again. "They can actually control exactly the number of electrons going into this capacitor," says physicist Per Delsing of Chalmers University of Technology in Göteborg, Sweden. On the face of it, what Zimmerman calls an electron "bucket brigade" is "very attractive as a fundamental current standard," he says.

At the moment, the pump generates currents at best 1000 times too small to be directly useful. But Zimmerman says the electron counter can be used to set a standard for capacitance, the ability of a device to hold electric charge. Zimmerman and Keller also hope to use the pump to test Ohm's law, a fundamental relation that states that the voltage across an electrical device is equal to the product of the current through it and its resistance. Using two tried and trusted methods, the Josephson standard for measuring voltage and the quantized Hall standard for resistance, to supply two of the three sides of the V = iR triangle, NIST's electron pump will supply standardized current to check the law for internal consistency.

The heart of the electron pump is a superconductor device known as a single-electron tunneling (SET) device. It exploits quantum tunneling, the ability of a particle—in this instance an electron—to leak through what is classically an impenetrable barrier. Electrons leak from the source electrode through one tunnel junction and into a conducting "island" beyond. Electrons can then escape the island via a second tunnel junction into the drain electrode. Voltage applied to a third electrode—the gate—adjoining the island influences the island region and allows electron flow from source to drain. As the gate voltage cycles up and down, electrons are sucked from the source and squeezed into the drain one at a time.



Seeing the light. NPL's single photon counter relies on superconducting magnetic detectors.

That's the theory. Because of quantum uncertainty, however, occasionally two electrons move through, and sometimes none. To get reliable electron counting, researchers must strap together several islands to form a pump. The simplest possible pump comprises three junctions separated by two islands, with each island handing off electrons to the next controlled by its own gate. The trick is to keep the gate voltages out of step to maintain the flow. The NIST pump contains seven junctions, enough to guarantee just one error for every hundred million electrons pumped through.

Chalmers's Delsing has also been working with several European metrology labs on developing ways of using an SET device to watch electrons as they pass through a circuit, rather than pumping them through. Their device works by passing the electrons through a cascade of islands and junctions, a little like the pump, while a separate SET transistor coupled to the chain sniffs electrons as they pass by. This electron counter is phenomenally sensitive, Delsing says. "The single-electron transistor is the most charge-sensitive device ever demonstrated," he adds.

Feeling the heat

In neighboring Finland, SET devices are already being used to measure temperature, exploiting the fact that the electron-trafficking of a string of junctions has a characteristic temperature behavior: a dip in its conductance at about zero voltage changes with temperature. The SET-based thermometer works only below about 30 K, but this could be its strength: "Below 0.65 K or so, there is no standard for temperature," Delsing says. "This might be a way to actually define the temperature scale at really low temperatures."

Another particle that can also be used to measure temperature is the phonon, a quantized ripple that flows through solids carrying heat from one place to another. The warmer the solid, the greater the population of phonons zipping around inside it: A grain of table salt at room temperature is home to about 10^{18} of them.

At NPL, Josephs-Franks and his colleagues Ling Hao and John Gallop are using a minute carbon nanotube as a single-phonon bridge between two objects at different temperatures. They have attached a nanotube, about one nanometer in diameter, to the tip of an atomic force microscope probe, forming a kind of proboscis. When they touch this on a surface, phonons can zip up and down the tube—as annular bulges in its wall—between the probe and the substrate. Instead of a smooth variation in temperature between the probe and substrate, the temperature changes in steps, says Josephs-Franks. The device could be used as a nanoscale thermometer to

CREDIT: PATRICK JOSEPHS-FRANKS/QUANTUM DETECTION GROUP

Published by AAAS

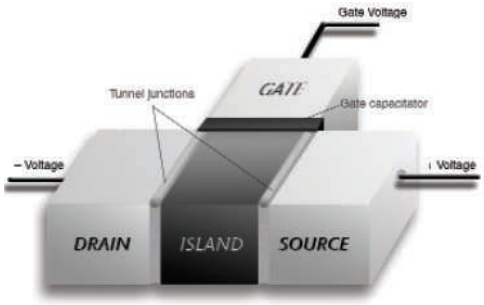
measure "temperature" at different points in macromolecules.

Counting light

Researchers have a number of tools, such as photomultipliers and avalanche photodiodes, that help them count light by the photon. But such counters don't meet the exacting demands of metrologists because they can't record each and every photon, and the range of frequencies they can see is limited. The dawn of the quantum information and communications era has given new impetus to the search for perfect photon detectors.

One such is a new superconductor-based photon counter, which relies on quantum tunneling, developed by Norman Booth at the University of Oxford and his collaborators at Harvard and the University of Naples. When an incoming photon strikes a superconducting surface layer, electron pairs are broken up, sending electrons tunneling into a second layer. Here, these electrons spawn yet more electrons, yielding a detectable electronic blip via a third layer. The device works from the infrared to the x-ray parts of the spectrum.

Sae Woo Nam of NIST Boulder and his colleagues have also developed a new detector for infrared photons using a thin film of superconducting tungsten. When an incoming photon heats the film, the tungsten momentarily loses its superconducting prowess and becomes an ordinary conductor. Researchers can spot the change by a sudden rise in electrical resistance. The tungsten returns to the superconducting state fast enough that the device can register 20,000 photons every second.



Go with the flow. The voltage on the gate in this SET transistor coaxes single electrons from source to island to drain through the seemingly impenetrable tunnel junctions.

At NPL, Josephs-Franks and his colleagues have tried a different approach to counting light. They use highly sensitive magnetic detectors called superconducting quantum interference devices (SQUIDs) to encircle a speck of photon-absorbing material. When a photon hits the absorber, it warms it slightly, causing a change in its magnetic properties that is detectable by the SQUIDs. One of the device's potential strengths is an ability to see many frequencies. "In principle it can work from infrared to x-ray," Josephs-Franks says. The trick is in finding the right absorber.

Although there's always scope for doing measurements better, the single particle seems to represent some kind of limit for metrology. The charge on the electron, for example, is believed to be constant over time and the same for every electron, so it makes for a perfect standard, says Zimmerman: "If charge comes in fundamental units and cannot be further subdivided, that seems to me like the end of the story."

–ANDREW WATSON

Andrew Watson, based in Norwich, U.K., is author of *The Quantum Quark*.

NEWS

Getting the Measure of Nanotechnology

In the realm of the very small, measurements lose their certainty and materials don't behave as they should. As chipmakers get to grips with circuits too small to measure, researchers are exploiting the oddities of the nanoworld to make new measuring devices

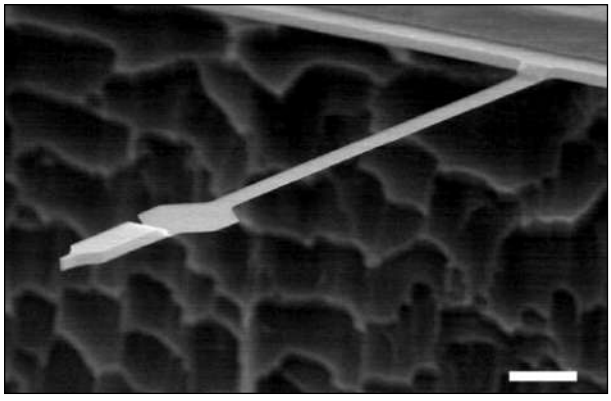
Next year, when the chip giant Intel ships its first Pentium 5 processors, the transistors inside will set a new record. Their gate length—the distance between source and drain—will have dwindled to 25 nanometers—just 100 atoms across. "That's indeed nanoelectronics in volume manufacturing," says Alain Diebold of International Sematech, an industry-backed R&D consortium in Austin, Texas. Even smaller things are on the way: Transistors with a gate length of just 5 nanometers have already been demonstrated in the lab.

For decades, semiconductor industry observers have warned that chipmakers will need new measuring and manufacturing technology to cope with shrinking chips. And for decades, engineers have managed to stave off the next technological revolution, as better electron microscopes and materials and clever optical tricks have made it possible to stencil finer and finer features onto semiconductor wafers. Industry has made a huge investment in its current instrumentation and so prefers to "squeeze a few more angstroms of resolution" out of it to keep it on the production line a little longer, says Michael Postek of the U.S. National Institute of Standards and Technology in Gaithersburg, Maryland.

Now, with 15-nanometer gate lengths looming and 10-nanometer lengths on the horizon, "we're going to begin to start feeling the pinch again," says Postek. "We know there's going to be a barrier coming. The metrology tools are beginning to fall off." This time, the experts say, radically different solutions may at last be unavoidable—and some are already in the works.

One reason the semiconductor industry may have to change its ways is that soon the minute circuits and transistors on chips will be sharing space with minuscule machines known as nanoelectromechanical systems (NEMS). A typical NEMS device incorporates a mechanical element sensitive to force and a means of converting mechanical energy into an electrical or optical signal. The other characteristic is size: NEMS features are typically 10 nanometers, or 40 atoms, across, offering extreme sensitivity to forces.

But designing and manufacturing NEMS brings new measurement challenges. At the scale of NEMS features, a distance that you measure may not be quite what you think it is. "Distance is not equal to distance," according to



Good vibrations. Oscillations in this minuscule diving board, which spans a human hair, can detect the magnetism of a single electron.

CREDITS (TOP TO BOTTOM): ILLUSTRATION: J. MOGLIA/*SCIENCE*; IBM

Published by AAAS

Jörg Kotthaus, a physicist at the Ludwig Maximilians University in Munich, Germany. With a ruler in the macroworld, there are clear endpoints, firm stops against which you can press the ruler. "In the nanoscopic world it's like trying to measure distance on a foam mattress," he says. And even if you can measure things accurately, two devices of the same size and shape may not behave exactly the same. Material properties rely on the assumption that a material is made up of a huge number of similar atoms or molecules. With NEMS, where the number of atoms or molecules is small, normal macroworld properties break down.

Device variability is something the semiconductor industry has learned to live with, and the outlook for NEMS devices is much the same, believes Harvard University's Robert Westervelt. "You have to do the engineering in such a way that that kind of stuff doesn't matter," he says. For example, at small scales the mass of a rod decreases much more rapidly than its linear dimensions because mass is linked to volume. As a result, a NEMS rod vibrates at a very high resonant frequency—up to 1 gigahertz, about the frequency of a mobile phone signal. But that tiny size also means that a few nanometers' difference in length, invisible irregularities in the material, or even the presence of a few contaminant atoms can dramatically change the rod's resonance response.

Despite the challenges, Kotthaus and his colleagues are using this type of tiny cantilever to make an "electron spoon" to transfer electrons mechanically from one spot to another, perhaps as a way of developing an electrical current standard. It's a scaled-down version of a classroom classic: Charge two parallel metal plates up to a few hundred volts and dangle a lightweight metallized ball between them—the ball will bounce back and forth, transferring charge. At Cornell University, Harold Craighead and colleagues have used a NEMS resonator in a different way, by measuring the change in resonant frequency of a tiny cantilever as a mass is added to the end. Craighead's device can measure masses as small as an attogram (10^{-18} g), and he is now developing it as a scale for weighing individual viruses and other biological structures.

Daniel Rugar and his colleagues at IBM's Almaden Research Center in San Jose, California, have also used a NEMS cantilever to create a sensor and now hold a record for the tiniest force detected by mechanical means. They used an ultrasoft cantilever 85 micrometers long and just 100 nanometers thick with a magnetic particle attached to the free end. When the cantilever is vibrated, the strong local field of the magnet can interact with a nearby single electron, which causes a small frequency shift in the cantilever. "We detected the magnetic signal from a single 'unpaired' electron spin," says Rugar. The technique is well-known in the macrolab, where it is called electron spin resonance, a sister of nuclear magnetic resonance. "Compared to conventional electron spin resonance, the technique is at least 10 million times more sensitive," says Rugar. "The ultimate goal is to develop a technique that can take three-dimensional [magnetic resonance] images of molecules with atomic resolution."

Ironically, although it is difficult to make precisely reproducible devices on nanometer scales, if you persevere, things get easier. "Once you go down to a certain scale, of course you have built-in rulers, the interatomic spacing," says Peter Cumpson of Britain's National Physical Laboratory in Teddington, near London. But although crystal lattices may provide a ruler, measurements of force become more murky. "It's difficult to separate your measuring apparatus from the device that you try to measure," Kotthaus says. "I can count atoms, but anything beyond that, as metrology, is difficult."

One way to tackle calibration issues, material matters, and length determinations is to press toward the quantum limit, says Westervelt. "If you start looking at single quantum particles, you get saved by the discreteness of the particles and counting," he says. Building the machines that can do this is incredibly hard, but once you climb that mountain, there are verdant pastures beyond. At the quantum limit, "it gets easier again," says Westervelt. Of course, you have to get there first.

–ANDREW WATSON

NEWS

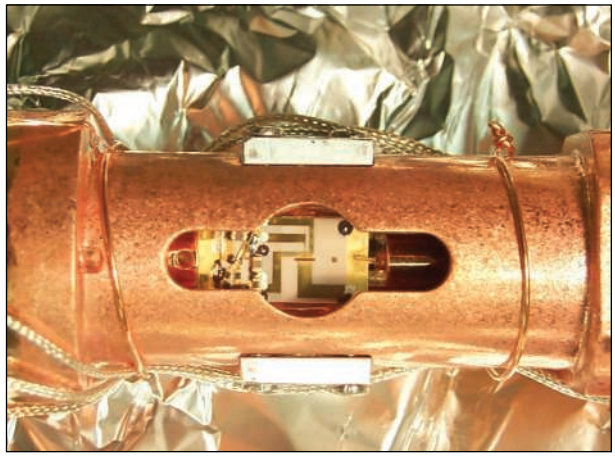
Time's Romance of the Decimal Point

Time's running out for the cesium atomic clock. To squeeze more decimal points of accuracy out of the world's top timekeepers, researchers are turning to some of the more esoteric corners of quantum mechanics

BOULDER, COLORADO—David Wineland perceives time much like the rest of us do: It races when you're trying to beat a deadline but crawls endlessly when you're stuck in traffic. But Wineland, a laser and atomic physicist at the U.S. National Institute of Standards and Technology (NIST) in Boulder, Colorado, has another relationship with time, one that most people simply can't fathom. He's working with his NIST colleagues to parse time more finely than anyone has ever achieved.

That's saying something, because the makers of the first atomic clock sliced a second so crisply that the uncertainty in their measurement was a mere 0.0000000001 of a second. That was in 1949. Today's record holder, NIST's F1 cesium atomic clock, has whittled that down to 0.0000000000000007 of a second. If a clock with that level of accuracy had been set up when the dinosaurs became extinct 65 million years ago, it would have dropped little more than a second by now. Yet, if the NIST team is successful, their clock—a mixture of lasers and trapped ions—wouldn't drop a second in 30 billion years, assuming anyone is still around to check.

This level of precision may seem esoteric; frankly, it is for most uses. But atomic clocks are now indispensable for a wide array of applications from satellite-based navigation systems and computer networks to managing the electrical grid and



Time trap. In this ion trap, the NIST team hopes to get aluminum and beryllium ions dancing in step.

CREDIT: D. WINELAND

Published by AAAS

cell phone traffic. The NIST researchers and others continue to push the technology, not just in the hope of launching novel applications that invariably stem from fundamental advances in measuring time, but also because they can. As Barry Taylor, a retired NIST physicist, was fond of saying, there is no field with a greater romance for the decimal point than the measurement of time.

All clocks rely on two basic components: a pendulum or some other "oscillator" that produces a regular set of "ticks," and a way to count and display those ticks as the passage of time. The trick to accurate timekeeping and pushing back that decimal point is reducing the outside influences on the clock's components. Small changes in temperature, humidity, and local gravity, for example, can change the movement of a pendulum or the grinding of gears enough so that over weeks and months a mechanical clock's precision slowly veers.

In 1945, U.S. physicist Isador Rabi suggested that the quantum-mechanical behavior of atoms could provide an oscillator that was largely immune to outside influences. Rabi noted that quantum mechanics confines electrons around atoms to certain energy states, and that hitting atoms with just the right frequency of electromagnetic radiation can cause electrons to jump from one state to another. In essence, atoms act like ultraprecise radio receivers that can tune in to a station at just one very specific frequency.

In cesium there is one such jump, called a hyperfine transition, that occurs when the atom is hit with microwaves with a frequency of 9,192,631,770 oscillations per second. And after cesium atoms undergo this transition, they fluoresce when hit with a precisely tuned laser pulse. So to use cesium as an atomic clock, researchers scan a sample of cesium atoms with microwaves across a range of frequencies until they see it fluoresce; then they hold their microwave emitter at that precise frequency. The oscillations of the microwaves are then the ticks of the clock, and researchers use other electronic detectors to count them.

Although cesium-based atomic clocks have been around for decades, clockmakers have continued to improve them, reducing error rates by about a factor of 10 every decade. Atomic clockmakers are also looking to new designs (see Diddams Review on p.1318 and Margolis Report on p.1355). Many hope to harness atomic transitions in cesium and other atoms triggered by optical lasers, which oscillate at frequencies as much as 1000 times that of microwaves and thus can dice the second into far more ticks. But one downside is that the laser-based traps that confine cesium atoms can interfere with the measurements. As a result, researchers must essentially fire the lasers through a measurement apparatus, limiting the time they can efficiently use their atoms.

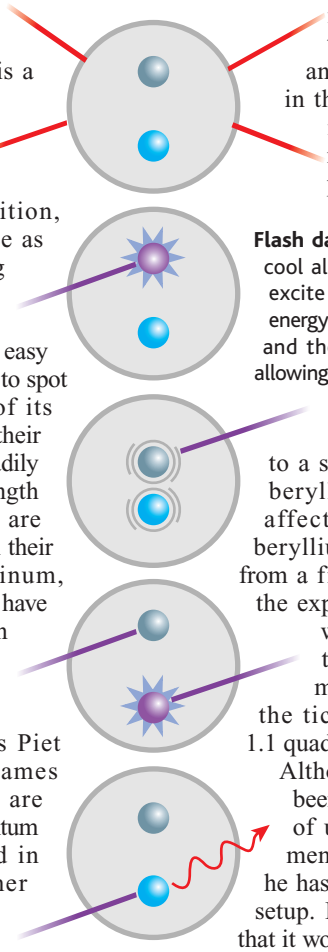
Some atomic clockmakers are working to get around this problem by trapping small numbers of ions in electromagnetic holding pens that don't interfere with the measurements, although this only works for some kinds of ions. The NIST group has been experimenting with aluminum, which works well in the electromagnetic traps but is a poor absorber of photons from lasers commonly used to cool atomic gases to the slow-moving state needed for accurate clocks. In addition, aluminum ions don't fluoresce as readily as other ions, making their transition difficult to spot.

But beryllium, it turns out, solves both these problems. It's easy to cool with lasers, and it's easy to spot the different energy states of its electrons: Beryllium atoms in their unexcited or "ground" state readily scatter photons with a wavelength of 313 nanometers, but they are transparent to this light when in their excited state. Unlike aluminum, however, beryllium atoms don't have an ultrasharp transition at an optical frequency from one energy state to another—bad news for an atomic clock.

So Wineland's colleagues Piet Schmidt, Till Rosenband, James Bergquist, and Wayne Itano are borrowing a trick known as quantum entanglement, which is used in quantum computing—another specialty of the Wineland lab—to link the states of the two ions so they can take advantage of the favorable attributes of each. Here, the goal is to transfer the atomic-transition information from the good clock atom, aluminum, to the good detection atom, beryllium.

The researchers start by trapping both an aluminum and a beryllium ion in a tiny chamber. These ions are further confined by a electromagnetic field that pushes them back toward the center every time they try to escape. Next, they use standard laser cooling methods to chill the ions to a fraction of a degree above absolute zero. Even though the aluminum ion doesn't readily absorb laser light to slow it down, it continually bumps into the beryllium ion, transferring some of its kinetic energy. And the beryllium ion, acting as a refrigerator for the aluminum, is then cooled with the laser.

Once trapped and cold, the ions are blasted with another laser, firing photons with a 267-nanometer wavelength. This time it's the aluminum ion's turn to go to work. The absorbed light kicks its electrons into a quantum-mechanical no-man's land called a superposition of states: part ground state and part excited state. Another laser pulse knocks aluminum back to its ground state and transfers its excitation energy into a change in motion of the combined aluminum and beryllium ions, which in their ultracold state move in concert like a single molecule. Next, another pair of laser pulses transfers this energy of motion to a superposition state of the beryllium ion, which in turn affects the fluorescence of beryllium in response to pulses from a final detection laser. Once the experiment is tuned up and working, the group uses the oscillation of the aluminum excitation laser as the ticking of the clock with 1.1 quadrillion ticks per second.



Flash dance. A series of laser pulses cool aluminum and beryllium ions, excite the aluminum, transfer this energy first to the motion of the ions and then to beryllium's electrons, allowing it to fluoresce.

Although Wineland says he's been kicking around the idea of using quantum entanglement in clocks for 10 years, he has only recently created the setup. Early indications suggest that it works, and the NIST team is working to nail down their results and tweak their laser setup to find the best clocklike energy transition in aluminum.

If the early hints hold up, it would certainly be a significant step, says Kurt Gibble, an atomic physicist at Pennsylvania State University, University Park. For now, however, he says it's still too early to tell whether this approach or others will ultimately push the decimal place the furthest. "The payoff is that the accuracy and stability should be record breaking," Wineland says. "We've been throwing around numbers like 10^{-18}," which corresponds to measuring time in quintillionths of a second. "But talk is cheap. We haven't done that yet," Wineland adds. If they do, it will create a new standard for measuring time that will not only be hard to beat but hard to fathom as well.

–ROBERT F. SERVICE

ILLUSTRATION: J. MOGLIA/*SCIENCE*

Published by AAAS

NEWS

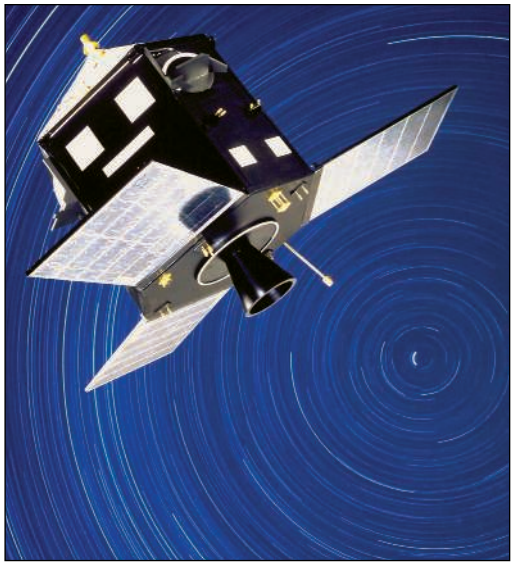
Putting the Stars in Their Places

Europe's Hipparcos set out to fix the positions of the stars with never-before-achieved accuracy. It did for most stars, but pinning down the distance of the Seven Sisters proved something of a headache

Astrometry, the science of measuring the positions of stars, has come a long way since Ptolemy. The Egyptian astronomer cataloged the positions of 1022 stars with a maximum accuracy of one-sixth of a degree. Today's best star catalog, compiled by the European Space Agency's (ESA's) Hipparcos satellite, lists the positions of a million stars, some to milli–arc-second accuracy. But Hipparcos has also sown confusion and uncertainty since astronomers discovered that its measurement of the distance to the best known star cluster in the sky—the Seven Sisters, or Pleiades—differs from their conventional distance estimate by 10%. The discrepancy could have repercussions for measuring the size of the universe.

Distance measurements are notoriously difficult in astronomy. Astronomers can get an accurate fix on stars within about 200 light-years of Earth using parallax: an apparent shift in the position of a nearby star against the unmoving background of distant stars, caused by Earth's motion around the sun. But the tiny parallax shift—usually a minute fraction of an arc second—becomes unmeasurably small for more distant stars. Parallax measurements show that a cluster known as the Hyades is 151 light-years away. On the assumption that stars in the Hyades and Pleiades are similar, 20th century astronomers compared their brightness and colors and calculated that the Seven Sisters are 440 light-years from Earth. Ever since, this distance estimate has been one of the steppingstones in determining the scale of the universe.

Enter Hipparcos, a revolutionary $300 million astrometry satellite launched by ESA in 1989. Using two precision telescopes, the rotating satellite mapped out a very accurate measurement grid on the sky from which it could derive stellar positions. Hipparcos plotted the positions of 1 million stars to a precision 15,000 times as accurate as Ptolemy's. For 118,000 stars, the accuracy was boosted another 20 times to milli–arc-second level, and precise parallax distances could be derived for stars out to hundreds of light-years. The resulting catalog, published in 1997, is the most accurate database of stellar positions ever produced. "Hipparcos put a lot of order into the slightly chaotic field of astrometry," says Hipparcos project scientist Michael Perryman of ESTEC, ESA's R&D facility at Noordwijk, the Netherlands.



Galactic positioning system. Hipparcos used two telescopes to map a precise grid onto the sky.

But while compiling the catalog, Hipparcos team member and star cluster expert Floor van Leeuwen, of Britain's Cambridge University, found that Hipparcos's parallax distance of the Pleiades was at most 400 light-years—10% smaller than the accepted value. The team had no reason to doubt the satellite's remarkable result. After all, explains Perryman, parallax is a purely geometric effect, whereas the older distance estimates relied on the astrophysical assumption that the Pleiades and the Hyades are similar. "Before you throw your hands up in the air and say that Hipparcos is wrong, you'd better be careful," he says.

Astrophysicists, however, were confident of their understanding of the Pleiades and instead cast doubt on Hipparcos. Early this year in *Nature*, for instance, astrophysicist Bohdan Paczyński of Princeton University suggested that the problem could be due to Hipparcos's highly eccentric orbit, the unintended result of a problem during the launch. In the same issue, Xiaopei Pan and colleagues at the California Institute of Technology in Pasadena presented new evidence from the Palomar Testbed Interferometer that one of the brightest Pleiades, called Atlas, is indeed 440 light-years distant. Perryman charges that jealous American astronomers are maligning Europe's groundbreaking satellite. Paczyński's remarks were "a comment given on the basis of ignorance," Perryman says. As for the Caltech team, he counters, "there's no guarantee that Atlas is at the center of the cluster, as Pan and his colleagues assume."

In a paper accepted for publication in *Astronomy & Astrophysics*, however, a team led by Susan M. Percival at Liverpool John Moores University in the U.K. convincingly shows that Hipparcos's mean distance to the Pleiades is indeed in error. Using visible and infrared observations of the cluster's stars, Percival and colleagues prove that the colors, luminosities, and chemical makeup of the stars make sense together only at a distance of 436 light-years. "It seems rather convincing," Perryman concedes.

Van Leeuwen now believes he may have found the reason for Hipparcos's error. The Pleiades constitute a "weak spot in the catalog," he says. He thinks the strong concentration of bright stars in the cluster threw off Hipparcos's delicate surveying strategy, resulting in a systematic error on the order of 1 milli–arc second. "There may be 10 or 20 other weak spots in the catalog," he says. "Most of them are other star clusters."

Van Leeuwen is now meticulously reanalyzing the huge Hipparcos data set and hopes to publish a new catalog sometime in 2005. "I haven't a clue yet what the new distance value for the Pleiades will be," he says. Perryman stresses that overall the Hipparcos results are still very reliable. "We'll end up with a catalog almost exactly the same," he says. "We're talking about extremely small, very local effects."

In 2009 and 2011, two new astrometry missions will be launched: NASA's Space Interferometry Mission and ESA's Gaia. Both are designed to provide much higher precision than Hipparcos, in the micro–arc-second range. Van Leeuwen says that important lessons have been learned from Hipparcos: Understanding and solving the Pleiades problem is of paramount importance for the data-analysis routines for Gaia. "No one ever realized that the system was so extremely sensitive."

–GOVERT SCHILLING

Govert Schilling is a writer in Amersfoort, the Netherlands.

CREDIT: ESA

Published by AAAS

NEWS

In the Blink of an Eye

Researchers want to freeze-frame the workings of atoms with laser pulses just billionths of a billionth of a second long. But first they must prove they really can produce a blast that short

To photograph something that happens very quickly, you need a camera with lightning-fast shutter speed. But to study how molecules behave and interact, no shutter is fast enough. Instead, for a couple of decades, researchers have used flashes of laser light little longer than a femtosecond; that's just a millionth of a billionth of a second. Now they want to go even quicker. Over the past few years, scientists have passed the femtosecond frontier and are measuring their pulses in attoseconds: billionths of a billionth of a second.

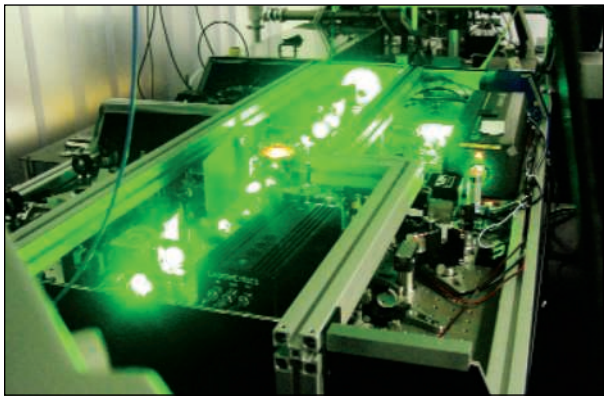
Such flashes should allow researchers to see inside an atom by freeze-framing the motion of an electron around the nucleus. They haven't got there yet—researchers are still learning how to produce the laser pulses cleanly and to measure their length—but they are looking forward to their first snapshots of the atom's interior. "I hope we will discover something we haven't even dreamed of," says Ursula Keller of the Swiss Federal Institute of Technology in Zurich.

A host of phenomena in molecules, atoms, and even solid matter takes place at attosecond time scales. "The dynamics of the electrons [during ionization] is much faster than a femtosecond. At one point they have to decide on which ion they go and sit, and this happens very fast," says Keller. After a decade of work, researchers are only just getting a glimpse of such processes. "At the moment there are very few systems that are able to make these measurements," says Ian Walmsley of the University of Oxford.

It's impossible to make an attosecond pulse with visible light because its wavelength lasts more than a femtosecond, and so the pulse would be less than a wavelength long. But in the early 1990s, several researchers suggested a way to make short pulses using shorter wavelengths, in the extreme ultraviolet (XUV) ranges. The technique, known as high-order harmonics generation, involves hitting atoms of a rare gas with a powerful femtosecond pulse from an infrared laser. As the electric field component of the infrared pulse oscillates back and forth, it rips electrons off the atoms, and then it smashes them back into the nucleus. As the electrons return to the ground state, they emit a burst of radiation that is a combination of higher harmonics of the applied infrared frequency. The result is a sharp attosecond-long XUV pulse.

Anne L'Huillier, now at Lund University in Sweden, pioneered this technique during the 1990s while working at the French Atomic Energy Commission's Saclay research center at Gif-sur-Yvette. But at first, researchers were only able to make strings of attosecond pulses about 1.3 femtoseconds apart. To get a snapshot of events inside the atom, they needed clean, isolated attosecond pulses. Part of the problem was that the infrared pulses used to make the attosecond flashes were themselves untidy and chaotic. The shape of the pulses—how the amplitude of the radiation rose to a peak then subsided again—bore no relation to the electromagnetic waveform that oscillated within it.

Researchers needed infrared pulses in which the maximum of the electromagnetic wave coincides with the maximum of the pulse envelope. Only that peak electromagnetic wave has the intensity to generate an XUV burst. In 1999, Keller's group proposed a way to make such a wave using a feedback mechanism that detects the state of the electromagnetic wave and tweaks the laser that produces it. But it was Ferenc Krausz of the Max Planck Institute for Quantum Optics (MPQ) in Garching, Germany, who turned theory into reality. In 2003, while he was at the Technical University of Vienna, his group reported neat single XUV pulses. "The Vienna-MPQ group is now clearly the leading group in this area. They have a system that works, and it works well," says Walmsley.



CREDIT: FERENC KRAUSZ/MAX PLANCK INSTITUTE

The light fantastic. Superfast laser pulses could open a new window on the workings of matter.

Although the pulses were undoubtedly short, Krausz and his team still had to prove that they were less than a femtosecond long. Earlier this year Krausz employed a technique known as a "streak camera" to measure the pulse length. He and his colleagues directed an XUV flash at a target of neon atoms. The pulse tears electrons from these atoms, and then the electric field of a second, infrared light pulse sweeps them sideways into an electron detector. From the energy distribution of these electrons, the researchers could determine the duration of the x-ray pulse—a speedy 250 attoseconds.

To demonstrate what attosecond pulses can do, Krausz and his team used them to make a waveform of light visible (*Science*, 27 August, p. 1267). In a technique they've dubbed the "light oscilloscope," the team ejected electrons from some atoms by blasting them with an attosecond XUV pulse and then hit those electrons with a femtosecond infrared pulse. During the small time window of 250 attoseconds, the electric field associated with the infrared light wave accelerates these electrons, which are then captured by a detector. From their arrival times and energies, the team could deduce the shape of the infrared light wave.

Several groups in Europe, North America, and Japan are now gearing up to do similar research, says Walmsley. "The tools and techniques of attosecond metrology are now really ready," says Krausz. One such team is led by theoretical physicist Thomas Brabec of the University of Ottawa. "We are working on potential applications in atoms and clusters … because [clusters] are the transition between atoms and condensed matter," he says. And Ahmed Zewail of the California Institute of Technology in Pasadena, who received the 1999 Nobel Prize in chemistry for his pioneering work in femtochemistry, is now also looking through this new window at matter in a state never seen before: "If you can catch any system in a very short time, then you are far from the equilibrium state of these systems, and in refining more and more the time resolution, you will find some interesting phenomena," says Zewail.

–ALEXANDER HELLEMANS

Alexander Hellemans is a writer in Naples, Italy.

Published by AAAS

VIEWPOINT

Metrology and the State: Science, Revenue, and Commerce

William J. Ashworth

"Natural measures of quantity, such as fathoms, cubits, inches, taken from the proportion of the human body, were once in use with every nation," taught Adam Smith in his lecture "Money as the measure of value and medium of exchange," delivered in 1763. "But by a little observation," he continued, "they found that one man's arm was longer or shorter than another's, and that one was not to be compared with the other; and therefore wise men who attended to these things would endeavour to fix upon some more accurate measure, that equal quantities might be of equal values. Their method became absolutely necessary when people came to deal in many commodities, and in great quantities of them (*1*)." Smith's comments and the rationale underpinning them became increasingly urgent toward the end of the eighteenth century.

The actual term "metrology," to describe weights and measures, was coined in the early nineteenth century by the mathematical examiner at Trinity House and the bookkeeping authority and ex–mathematical master at Finsbury Square Mercantile School, Patrick Kelly. The requirements of increased trade and the fiscal demands of the state fuelled the march toward a regular form of metrology. Measures originally gained their meaning (and practice of gauging) from the local understanding of the object being measured. For an emerging integrated national market to properly function, a reduction in the number of different types and versions of weights, measures and containers is required.

A uniform system of taxation meant accounting for foreign and domestic customary variations in weights, measures, and containers. Not surprisingly this could be a tiresome, complicated, and time-consuming process. As a result, the state's revenue activities gradually impinged upon the diversity of British and colonial metrological practices and containers and packaging, because it tried to recast such things to aid its own activities. This was not without immense opposition. Such a preoccupation was also the obsession of other European countries during this period. Measurement as such was not the primary issue; rather, it was the fact of a state-defined version, increasingly alienated from the object being gauged, being implemented over local versions that really rattled dispersed communities. The state, after all, was hardly the most trusted agglomeration of institutions, with the board of excise quite literally the least. Everyday folk may have been suspicious of state approaches to quantification, but they themselves lived by their own version, dominated by a local notion of a "just measure." To have transregional (let alone international) standardized abstract measures requires a legitimating form of knowledge, the agencies to enforce it, and a process of regional education (*2*–*4*). The interesting issue is not diversity but rather when diversity was seen to be a problem.

School of History, University of Liverpool, Liverpool L69 3BX, UK. E-mail: W.J.Ashworth@liverpool.ac.uk

Diversity of Meaning

Legislating for a system of regularized weights and measures was one thing, but making containers to strict specifications, that is, all the same, was another. The technology, skills, and sheer cost in manufacturing standardized casks (or packaging in general) was simply not feasible. This problem haunted all eighteenth-century attempts of imposing accurate measures like that of the bushel. How could a village Turner or Cooper correctly calculate and build an accurate representative of a bushel? What materials should be used, and what should be the relation between circumference and outer body? How could the vessel be made to avoid tampering, and how should the grain be poured into the vessel?

The heavily taxed item of coal is a case in point. It was generally sold by volume rather than by weight, and this depended upon capacity. The vessels used to measure the coal varied greatly from place to place and over time. For example, the Newcastle chaldron increased by a factor of 3 over the course of 150 years. W. D. Patlenden (*5*) claimed the changes in measures in the coal trade were because of technical developments and tax evasion. By far the largest consumer of coal was London, which received its supply from the northeast and especially Newcastle. Each shipload of coal was levied according to the number of chaldrons or keels it carried. The commissioners examining the public accounts during the 1780s complained that a chaldron was ''different at different Places. The chaldron at the Port of Lading, whether Newcastle or Sunderland, is more than the chaldron at the Port of London (which is according to the Winchester Measure) in the proportion nearly of Twenty-one to Eleven (*6*).'' The actual gauging was done by men known as ''meters'' who were appointed by the commissioners of customs. Not only could the size of chaldrons vary, but their value depended on the size of the pieces of coal and their water content. Merchants would buy their coal in lumps as large as possible and sell them in medium sizes known as ''round coal.'' This was abolished in the Weights and Measures Act of 1835, which legislated that from January 1836 all coal was to be sold by weight only. Similar problems plagued the gauging and selling of an array of other items, including grain and salt (*5*–*7*).

One problem in trying to establish standards was the fact that official state institutions holding original weights and measures differed from one another. The exchequer's standards stemmed from the reign of Henry VII and was frequently the one named in legislation. A statute passed during his reign legislated that ''standard weights and measures be made and sent to the several Cities, Boroughs, and Market Towns therein mentioned.'' This was subsequently done, but soon ''the said standard weight and measures were found defective.'' Consequently, another statute was passed in which it was specified that ''the measures of a Bushel shall contain eight Gallons of wheat, and that every Gallon contain eight Pounds Troy of Wheat, and that every Pound contain 12 ounces Troy weight, and every Ounce contain 20 sterlings, (now 20 Penny-weight,) and every Sterling, or Penny-weight, be of the weight of 32 corns of wheat that grew in the Middle of the Ear of Wheat, and that a Standard of a Bushel and a Gallon after the Assize be made and kept in the King's Treasury for ever.'' The new measures were thus ordered and distributed while the old ones were returned and destroyed. As a result, the exchequer now contained a standard brass bushel and a standard gallon (*8*).

In February 1696, during the passage of an extremely important bill concerned with establishing an excise duty on malt, an experiment was conducted in the presence

of certain members of Parliament and several excise officers—George Tollet, Philip Shales, Thomas Jeff and probably the most authoritative economic technician and gauger of the time, Thomas Everard—to ascertain the content of the standard bushel. It was decided that the said standard should be a cylindrical vessel with a diameter of 18.5 inches, a depth of 8 inches, and contents of 2150.42 solid inches. The dimensions were rounded off to these figures to make them ''convenient… without counting to the hundredth Part of an Inch.'' This new standard for the Winchester bushel was made law in the above malt act. The excise officers also compared the standard troy weights with the standard avoirdupois weights and found that 15 pounds avoirdupois was equal to 18 pounds, 2 ounces, and 15 pennyweight troy. Hence 140 ounces avoirdupois was equal to 218.75 ounces troy. Therefore ''The Bushel, as now settled, contains 2150.42 Solid Inches…and will contain of common spring water 1134.344 Ounces Troy'' (*9*, *10*).

The measure for wine was taken from the gallon sealed at the Guildhall in London. It was officially by this measure that all wines, brandies, spirits, strong waters, mead, perry, cider, vinegar, oil, and honey were to be measured and sold. The Guildhall wine gallon was assumed to contain 231 cubic inches, whereas a hogshead was presumed to hold 63 gallons. However, following a claim made by a certain Dr. Wybard that the standard wine gallon actually only contained 224 to 225 cubic inches at the most, two general excise officers, Richard Walker and Philip Shales, made an experiment to test his claim. They carefully constructed a vessel out of brass in the form of a parallelepipedon with sides 4 inches long and a depth of 14 inches. This gave a volume of 224 cubic inches. They presented the vessel at the Guildhall in London on 25 May 1688 to an audience consisting of the lord mayor; the commissioners of the excise; the astronomer royal, John Flamsteed; the Oxford astronomer Edmund Halley; and several others. The vessel was filled with water and emptied into the old standard wine gallon, which was filled exactly. Nevertheless, ''for several reasons, it was at that time thought convenient to continue the former supposed content of 231 Cubic Inches to be the Wine Gallon, and that all Computations in gauging should be made from thence as above.'' Thus, there was no check on the gauger through a reliable standard vessel (*11*, *12*).

The gallon for wine had been defined several centuries earlier, in 1303, as a vessel containing eight tower pounds of wheat. All the subsequent legislation referred to wine gallons, which, without specifying, thus legally stemmed from this definition. The story gets even more confusing. Lord Crayford's later committee on weights and measures (1758) was curious to ascertain why customs and excise gaugers used a wine gallon of 231 cubic inches, which was 51 cubic inches less than the beer and ale gallon of 282 cubic inches. The commissioners of excise told them that they believed the difference stemmed from a memorial dated May 1688 from the commissioners of excise and hearth money to the treasury. It seems that there was some initial confusion after the results confirming the Guildhall experiment, showing the wine gallon to be 224 cubic inches. Originally the excise commissioners recommended the standard to be taken at this revised figure. Merchants soon got wind of the suggested change and enquired whether they could sell at the new standard. However, the powerful attorney general, Sir Thomas Powys, quickly ruled against any alteration because it would adversely affect the revenue (*13*). Precision and accuracy (in the sense of obtaining a constant result) was invariably a factor of legislation, commercial procedures of convention, and, vitally (as revealed here), the crown's purse.

Metrology and the State

Between 1660 and 1714, acts were passed that attempted to define the measures to be used nationwide for ale, beer, coal, corn, herrings, soap, salt, fruit, malt, cider, and perry. These were mainly excised goods clearly demonstrating that revenue concerns were one of the motivating factors. The Winchester bushel was imposed upon the malt trade in a financial act of 1701, accompanied by the claim that ''there is a great variety of Bushels and other Measures of different Contents and Gauges used…for the measuring, buying and selling of all sorts of Graine, Salt and other Commodityes…to the great defrauding and oppressing of the people.'' The same sentiments appear to be behind an act the year before concerning measures for retailing ale and beer (*14*–*17*).

All of this poses a paradox. To appear fair, taxation should be universally applied and governed by a set of standards equitably applied. However, the imposition of such measures required illiberal methods that often rode roughshod through widespread diversity. Within this context, the work of the Polish historian of metrology, Witold Kula (*18*), is particularly useful. Kula has demonstrated that, before the establishment of the metric system on the continent over the course of the nineteenth century, concrete concepts such as the finger, foot, and ell (elbow) were in everyday use. They had no abstract, standardized denomination, and accounting for the weight or measure of a commodity was a qualitative process that varied from region to region (and indeed within regions). It was a process suited to small communities and local markets. Consequently, making measures accountable to a centralized source of social authority was extremely difficult.

Accompanying the growth of the state's power and the expansion of its reach and combined with increased commerce and expanding markets during the second half of the eighteenth century, weights and measures were increasingly made accountable to an abstract standard separated from people's everyday lives and work. This argument is neatly summarized by Theodore Porter: ''Informal measurement was inseparable from the fabric of these relatively autonomous communities. It broke down with the intrusion of more centralised forms of power—both political and economic—with the relatively private domain of communal life (*19*).'' The people that suffered were most frequently those excluded from some form of institutional power. As Peter Linebaugh powerfully showed (*16*) with regard to the Atlantic tobacco trade: ''The class struggle in the oceanic tobacco trade took a metrological form, because the ambiguities of measures benefited the porters, the crews, the slaves, the lightermen and the 'little inconsiderable persons.' Legislation attempted to standardise the hogshead.'' Greater regularization and centralization was accompanied by increased abstraction, the antithesis to localism and diversity. To legitimate this abstraction and make it appear real, an accompanying form of reason that appealed to a notion of objectivity (and equity) was required (*16*, *18*–*20*).

For much of England's history, the standard of length appropriately had its basis in the nation's most important source of food, the barleycorn. It had to be ''taken out of the middle of the ear, and being well dried, three of them in length were to make one inch; and thence the rest.'' Similarly, the standard weight derived from ''a corn of wheat gathered out of the middle of the ear: which being well dried, 32 of them were to make one penny-weight, 20 penny weights one ounce, and 12 ounces one pound troy'' (a total of 7680 grains). In actuality, there were as many as six different pounds. For example, another troy pound was used for gold and silver that weighed 5760 grains; the tower pound that was used to test coins weighed 5400 grains; and a wool pound, at 6992 grains, was used to weigh ordinary goods. Under Elizabeth I, an unsuccessful attempt was made to impose a single troy pound weighing in at 7000 grains for the purpose of gauging all ordinary items, whereas in 1758 Parliament decided to legalize the single troy pound and enforce the avoirdupois pound for weighing heavy goods. For measures of capacity, a unified approach was taken for both dry and liquid

goods: ''eight pounds troy weight of wheat, gathered out of the middle of the ear, and well dried, shall make one gallon of wine measure; and that there shall be but one measure for wine, ale, and corn, throughout this realm'' (*10*, *21*, *22*).

The historian Julian Hoppitt made the case (*14*) that much of the legislation passed under the later Stuarts concerning weights and measures could be interpreted as an attempt to bring geographically remote areas into line with the more economically active regions of the south and east. Certainly under the Act of Union with Scotland in 1707 an unsuccessful attempt was made to bring areas together through metrology. One of the articles of the act stipulated that Scotland had to adopt the legal weights and measures of England. The earl of Godolphin complained to the Scottish commissioners of excise that the English gallon in the exchequer was some 10 cubic inches less than the Scottish equivalent. In July 1707, the Scottish commissioners of customs in Scotland were quick to highlight the need for uniform weights and measures between the two newly unified nations: ''By reading the 17th Article of the Union, it occurs to us that we ought to have weights and measures of England sent here forthwith, at least patterns, to the end, every port and place be furnished; for we conceive that without these calculations cannot be made, and it is a great trouble to us that early care was not taken thereof.'' Abolishing the Scottish system of weights and measures simply meant Scotland used two systems. The Scottish courts defended the continued use of local measures for internal trade, whereas London directed customs and excise to collect duties with use of English measures. Trying to make a coherent and therefore predictable tax policy under these conditions was immensely difficult (*23*–*25*).

Attempts to enforce the Winchester bushel returned in 1732 under the auspices of Robert Walpole's administration; again the project failed. In 1742 an anonymous gentleman of the Royal Society was struck by the diversity of supposed standard weights and measures kept in the various London locations, which, as we have seen already, were meant to hold the original authoritive ones. A few years later Crayford's select committee report on weights and measures was published, followed by a second report the following year. Its contents centered upon the inadequacy of current legislation and the weak process of enforcement. Although attempts were made to act on the committee's resolutions, the bills that were subsequently introduced were so late in the session they failed (*8*, *26*).

Crayford's committee examined all the standards kept in the government's exchequer depositories at the Guildhall, Founders Hall, the Watchmakers Company, and the Tower of London. Among the members of the committee were the president of the Royal Society, Lord Macclesfield, and several prominent mathematicians and astronomers. They universally condemned the various official liquid gallon measures and advocated that the wine gallon kept at the exchequer, and not the one housed in the Guildhall, be adopted (*27*). The choice was guided by the quest to centralize all the measures and the fact that it was the most commonly used. Despite the committee's failure to pass any legislation, it did instigate new standards for the troy pound and yard, which were constructed in 1758 and 1760 by the mathematical instrument maker John Bird. These were subsequently made the primary references for the imperial system established in 1824. One problem mitigating the standardization and enforcement of weights and measures was the vast array of legislation that allowed exemptions. For instance ''one Act permitted Oats, Malt, and Meal, to be sold differently from other Corn; that was repealed after 20 Years Practice had habituated the People to that way of selling: Another Act excepted, the county of Lancaster, because in the county a larger measure was in use than the Law allowed; and many other instances of the like kind might be shewn (*28*).'' The locally informed and therefore haphazard nature of legislation until this point was thwarting what the Crayford committee termed ''the Principles of Uniformity.'' The committee concluded that ''in order effectually to ascertain and enforce uniform and certain Standards of Weight and Measures to be used for the future, that all the Statutes relating thereto should be reduced into one Act of Parliament; and all the said Statutes now in being, subsequent to the Great Charter, repealed.'' But it was precisely because communities were so devoted to local measures that members of Parliament couldn't agree on authorizing a system that overturned such a highly charged context. As late as 1817, it was estimated by the agriculturist expert on the distillation of spirits and minister of Keith Hall and Kinkell, George Keith, that throughout provincial England there were ''about two hundred and thirty'' different weights and measures and a further 70 in Scotland. He also added that it was extremely frequent to find several different weights and measures in the same county (*28*, *29*).

In 1814 a commons select committee reported that ''the great causes of the inaccuracies which have prevailed, are the want of a fixed standard in nature, with which the standards of measures might at all times be easily compared, the want of a simple mode of connecting the measures of length, with those of capacity and weight, and also the want of proper Tables of Equalisation, by means of which the old measures might have been made to establish a mode of connecting the Measures of capacity with weight.'' Nature had by now become the state's legitimating authority to crush localism. However, the problem was twofold: First, there was no physical standard with which to police deviation, and secondly, as previously seen, there was confusion generated by the proliferation of statutes concerned with variations in weights and measures. The 1814 committee was composed of 23 members, all now reliant upon scientific information supplied by two of Britain's leading men of science, the experimental natural philosopher and physician, William Hyde Wollaston, and the professor of natural philosophy at Edinburgh University, John Playfair, both of whom had worked extensively on pendulum vibrating seconds in the latitude of London. A bill was put forward in 1815 ''for establishing and preserving an uniformity of weights and measures'' but failed after its second reading. A new committee was subsequently formed in 1816 to further investigate pendulum-vibrating seconds as a source for grounding metrological standards, this time led by the member of Parliament for Bodmin and future president of the Royal Society of London, Davies Gilbert (*14*, *30*).

The new committee again contained the elite of British men of science, many of whom were, or went on, to be employed in major state institutions. The members consisted of the commissioner of the board of longitude (from 1818), Wollaston; the secretary of the board of longitude (from 1818), Thomas Young; the leading precision instrument maker and fellow of the Royal Society, Edward Troughton; the natural historian and president of the Royal Society, Joseph Banks; the secretary of the Royal Society and until 1814 a medical officer in the army, Charles Blagden; the army surveyor and fellow of the Royal Society, Henry Kater; the secretary of the Admiralty and one-time topographer of South Africa, John Barrow; and lastly, the lieutenant governor of Woolwich Military Academy, one-time director of the Ordinance Survey, and fellow of the Royal Society, William Mudge. Their results were brought out in 1818, but again their findings ultimately failed to come up with a solution. Yet another committee was subsequently formed on how best to define the standards and implement them into legislation. This time it was led by the former members of the above committee along with the addition of the minister of Parliament and, from 1819, lord of the Admiralty, George Clerk (*31*).

The work of these and additional committees finally culminated in the imperial system of weights and measures in 1824. The new metrology was aimed at reaching a balance between scientific objectives, practical requirements, and commercial reception. The implementation of abstraction to overcome localism and diversity was considered too dangerous to be made the evangelical basis of a new metrology. It had to be consistent, recognizable, and simple in the sense of being easily understood and enforceable. Exactness was a negotiation of all these boundaries. Unlike the French metric system, which was perceived in Britain as having been an expensive failure and commercial disaster, the 1824 solution was a pragmatic compromise. The customary practices and use of old measures were far too deeply ingrained to be simply replaced at a stroke, as had been demonstrated in France. As the precision instrument maker Jessie Ramsden had observed in 1792 while investigating the standard for proof spirit (*32*), ''To retain the present value of Proof, will, no doubt, have many advantages: it will prevent that confusion which always happens in commerce, when any change of value, or denomination, of merchandise takes place.'' The new imperial weights and measures took the most widespread and everyday consistent standards in use and simplified them into a coherent system. The key imperative of the act was to ensure as little disruption as possible to the commercial environment (*14*, *32*, *33*).

Lineal standards were now regularized and derived from the imperial standard yard, which was based on a pendulum vibrating seconds in London at the proportion of 36 to 39.1393. The standard for all measures of capacity derived from the imperial standard gallon, which contained 10 pounds weight avoirdupois of distilled water weighed in air at a temperature of 62° Fahrenheit and a barometer reading at 30 inches. All duties, allowances, drawbacks, payments, and accounts under any law of excise were to be made to these standards. This, of course, meant that all prior existing statutes related to this issue were repealed. For example, the wine gallon of 231 cubic inches was replaced by the imperial gallon, defined as 277.274 cubic inches and being the space taken up by water poured into a 10-pound avoirdupois weight at the legally defined temperature and pressure. The imperial standards established three weights and measures from which all other metrological standards derived. They were the yard, the troy pound, and the gallon. All measures of length were now to stem in parts or multiples of the yard, as constructed earlier by John Bird at the request of the Crayford committee in 1760. Similarly, all weights were now derived from the troy pound as originally constructed, once again, by Bird in 1758 (*34*, *35*).

The next really significant legislation came with the Weights and Measures Act of 1835, which legally abolished the Westminster bushel and all local and customary measures in the marketplace, along with practices such as striking the commodity. Everything now had to be sold by the imperial bushel. The combination of this and the earlier 1824 act inspired Joshua Bateman, the author of a very popular nineteenth-century excise manual (*36*), to gleefully declare ''The Winchester bushel, and all local or customary measures are abolished. Heaped measures are also abolished.'' To ensure that time was saved and mental labor was mechanical, all calculations were ''reduced to tables.'' In addition, inspectors were authorized to check all weights and measures in their own areas. ''Great Britain,'' declared a recent historian of metrology, ''was on the verge of creating one of the most efficient metrological officer corps in European history'' (*34*).

References and Notes

1. A. Smith, in *Lectures on Justice, Police, Revenue and Arms; Delivered in the University of Glasgow and Reported by a Student in 1763*, E. Cannan, Ed. (Oxford Univ. Press, Oxford, 1896), pp. 183.
2. W. J. Ashworth, *Customs and Excise: Trade, Production and Consumption in England 1640–1845* (Oxford Univ. Press, Oxford, 2003), chap. 15.
3. S. Schaffer, in *Rethinking Social History: English Society 1570–1920 and its Integration*, A. Wilson, Ed. (Manchester Univ. Press, Manchester, 1993), pp. 128–157.
4. For plebian suspicions of state quantification, see R. Sheldon, A. Randall, A. Charlesworth, D. Walsh, in *Markets, Market Culture and Popular Protest in Eighteenth-Century Britain and Ireland*, A. Randall, A. Charlesworth, Eds. (Univ. of Liverpool Press, Liverpool, UK, 1996), pp. 25–45.
5. W. D. Patlenden, in vol. 20 of *Bulletin of the Cleveland and Teesside Local History Society* (Society of Middlesbrough, 1973), pp. 12–14.
6. For the commissioners, see U.K. House of Commons, "Thirteenth report of the committee appointed to examine the public accounts," *Sessional Papers*, 1785, vol. 44, p. 18.
7. See also the Public Record Office, Customs 41/1, 14 July 1713 (Kew, UK, 1713).
8. C. Leadbetter, *The Royal Gauger; Or, Gauging made Easy, As it is actually practised by the Officers of His Majesty's Revenue of Excise* (E. Wicksteed, London, ed. 2, 1743), p. 119.
9. C. Leadbetter, *The Royal Gauger; Or, Gauging made Easy, As it is actually practised by the Officers of His Majesty's Revenue of Excise* (E. Wicksteed, London, ed. 2, 1743), p. 120.
10. One pound troy was in proportion to the cubic inches contained in a wine gallon, whereas 1 pound avoirdupois was "very near" the cubic inches contained in an ale gallon.
11. "Proposals for altering the gauge of wines and brandies," [Cholmondley Houghton Papers, Cambridge Univ. Library, manuscript no. 44/93, n.d. (probably early 1730s)].
12. C. Leadbetter, *The Royal Gauger; Or, Gauging made Easy, As it is actually practised by the Officers of His Majesty's Revenue of Excise* (E. Wicksteed, London, ed. 2, 1743), p. 117.
13. U.K. House of Commons, "Report from the committee appointed to enquire into the original standards of weights and measures in this kingdom," *Sessional Papers*, 1758, vol. 19, p. 376.
14. J. Hoppitt, *Eng. Hist. Rev.* **108**, 82 (1993).
15. R. E. Zupko, *Revolution in Measurement: Western European Weights and Measures Since the Age of Science* (Memoirs of the American Philosophical Society, Philadelphia, 1990), p. 26.
16. P. Linebaugh, *The London Hanged: Crime and Civil Society in the Eighteenth Century* (Penguin, London, 1991), pp. 162–163.
17. For the seventeenth-century commentator, see *A Sure Guide for His Majesties Justices of Peace* (Penguin, Harmondsworth, 1663), p. 441.
18. W. Kula, *Measures and Men*, R. Szreter, Trans. (Princeton Univ. Press, Princeton, NJ, 1986), p. 70.
19. T. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton Univ. Press, Princeton, NJ, 1995), p. 223.
20. Colonies usually, but not always, adopted the measures of the occupying country (*24*).
21. *Brit. Rev. London Crit. Rev.* **IX**, 159 (1817).
22. J. T. Graham, M. Stevenson, *Weights and Measures and Their Marks: A Guide to Collecting* (Shire Publications, Haverford-West, ed. 3, 1993), pp. 3–4.
23. G. S. Keith, *Different Methods of Establishing an Uniformity of Weights and Measures Stated and Compared* (J. Johnson & Co., London, 1817), p. 1.
24. J. J. McCusker, *William Mary Q.* **30**, 599 (1973).
25. For numerous other Scottish-English legal variations in metrology, see R. E. Zupko, *Revolution in Measurement: Western European Weights and Measures Since the Age of Science* (Memoirs of the American Philosophical Society, Philadelphia, 1990), pp. 5–7.
26. "An account of a comparison lately made by some gentlemen of the Royal Society of the standard of a yard, and the several weights lately made for their use," *Philos. Trans. R. Soc.* **42**, 541 (1742).
27. During the Middle Ages, City of London officials were in charge of stamping and testing bronze and brass weights at the Guildhall jointly with the Worshipful Company of Founders until 1579 (*37*).
28. U.K. House of Commons, "Report from the committee appointed to enquire into the original standards of weights and measures in this kingdom," *Sessional Papers*, 1758, vol. 19, p. 352.
29. G. S. Keith, *Different Methods of Establishing an Uniformity of Weights and Measures Stated and Compared* (J. Johnson & Co., London, 1817), p. 2.
30. R. E. Zupko, *Revolution in Measurement: Western European Weights and Measures Since the Age of Science* (Memoirs of the American Philosophical Society, Philadelphia, 1990), pp. 105–107.
31. R. E. Zupko, *Revolution in Measurement: Western European Weights and Measures Since the Age of Science* (Memoirs of the American Philosophical Society, Philadelphia, 1990), pp. 107–108.
32. W. J. Ashworth, *Technol. Culture* **42**, 27 (2001).
33. For an account of the implementation and reception to the French metric system, see K. Alder, in *Values of Precision*, M. Norton Wise, Ed. (Princeton Univ. Press, Princeton, NJ, 1997), pp. 39–71.
34. R. E. Zupko, *Revolution in Measurement: Western European Weights and Measures Since the Age of Science* (Memoirs of the American Philosophical Society, Philadelphia, 1990), pp. 178–179.
35. B. R. Leftwich, *Customs: The Story of a Great Department* (Customs and Excise Museum, Liverpool Galleries and Museums, n.d.), p. 207. John Bird gained immense prestige during the eighteenth century for the precision instruments he made for the astronomer royal, James Bradley.
36. J. Bateman, *The Excise Officer's Manual, and Improved Practical Gauger; Being a Compendious Introduction to the Business of Charging and Collecting the Duties of Excise* (A. Maxwell, London, 1840), pp. 28–29.
37. J. T. Graham, M. Stevenson, *Weights and Measures and Their Marks: A Guide to Collecting* (Shire Publications, Haverford-West, ed. 3, 1993), p. 1.

REVIEW

Standards of Time and Frequency at the Outset of the 21st Century

S. A. Diddams,* J. C. Bergquist, S. R. Jefferts, C. W. Oates

After 50 years of development, microwave atomic clocks based on cesium have achieved fractional uncertainties below 1 part in 10^{15}, a level unequaled in all of metrology. The past 5 years have seen the accelerated development of optical atomic clocks, which may enable even greater improvements in timekeeping. Time and frequency standards with various levels of performance are ubiquitous in our society, with applications in many technological fields as well as in the continued exploration of the frontiers of basic science. We review state-of-the-art atomic time and frequency standards and discuss some of their uses in science and technology.

As important as "time" might be to those who are navigators, scientists, or even musicians, it is no more than an arbitrary parameter that is used to describe dynamics, or the mechanics of motion. David Mermin was struck by this as he wondered about the role that time and space would play in physics in the next century (*1*):

> *How can people talk about spacetime turning into a foam at the Planck scale when we barely manage to define space and time at the atomic scale? Time, for example, is nothing more than an extremely convenient and compact way to characterize the correlations between objects we can use as clocks, and clocks tend to be macroscopic. To be sure, we can generate frequencies from atoms and correlate them with macroscopic clocks, but the shorter the length scale, the more it looks like you're talking about energies divided by Planck's constant. The connections with clocks become increasingly indirect. There seems to me to be a considerable danger here of imposing on an utterly alien realm a useful bookkeeping device we've merely invented for our own macroscopic convenience.*

The definition of time can be puzzling exactly because of the apparent arbitrariness that Mermin described. It is through the external or internal periodic dynamics of one object that we define time, and armed with that time scale, we can characterize the dynamics of other objects—an oddly circular argument. Another conundrum: How do we determine the period (or its inverse, frequency) of our time standard, or any clock, to be uniform? Clearly, time is relative and several time sources must be compared toward establishing the most stable and accurate definition of the second, our base unit of time in the international system (SI).

In view of this, there is considerable irony in the fact that the second is the most accurately realized unit of measurement, with fractional uncertainty now below 1 part in 10^{15}. Moreover, it could be argued that the technologies based on the arbitrarily defined second have had an impact like few others in our modern society. Everyday systems including the electric power grid, cell phones, the Internet, and the Global Positioning System (GPS) depend critically on time and frequency standards for continued operation. Because of its position of metrological preeminence, the second is also used to define three other SI units (meter, candela, and ampere), and several other important physical quantities are defined or measured in terms of the second. For example, accepting that the speed of light is a constant, the meter is defined as the path length traveled by light in a vacuum during the time interval of 1/299,792,458 of a second. As a result, lasers with a known and fixed frequency have become the standards for length metrology, where they guide precision measurements of physical distances (e.g., in the etching and lithography of semiconductor wafers). Another example is the definition of the volt, which can be obtained via the Josephson effect in terms of the product of a physical constant and the frequency. For these and other reasons, the development and operation of high-quality time and frequency standards is an important endeavor at National Metrology Institutes (NMI) around the world, with consequences that reach into our daily lives.

In many research laboratories, the scientific impact of precision time and frequency measurements has been considerable. Examples include the prediction of gravitational radiation from binary pulsars (*2*) and the most precise direct measurement of the gravitational red shift (*3*). Additionally, the development of time and frequency standards over the past 60 years has gone hand-in-hand with many scientific advances in the fields of atomic, molecular, and optical physics. Modern atomic clocks have their roots in the microwave spectroscopy experiments of Stern, Rabi, and Ramsey. Many of the techniques that are used in modern atomic clocks resulted from the development of the maser, the laser, and the new field of laser spectroscopy that followed. These developments led to laser cooling and trapping of atoms and ions, which now provide clockmakers with isolated and nearly motionless quantum references for what are now the best clocks in existence.

In labeling the time of a physical event, one must count the number of cycles (and perhaps fractions of cycles) of some periodic occurrence relative to an agreed-upon time origin. For some situations, the time origin is of great importance; in other cases, it is only the interval, or time difference, between two events that is of interest. In what follows, we focus on the latter and discuss the issues surrounding the generation and characterization of the source of the periodic events (often called a frequency standard) with which time intervals are generated and measured. As many excellent reviews on the development of atomic frequency standards already exist (*4*–*6*), we concentrate on the most accurate atomic clock—the cesium fountain clock—and then discuss the new optical clocks, which are anticipated to be the atomic timepieces of the future.

Historical Background

The best choice for a timekeeping device is an object whose dynamical period is well characterized, not easily perturbed, and (ideally) constant. A natural, macroscopic candidate that could be used to define the unit of time is some phenomenon of nature, whose period is especially uniform. Figure 1 compares the performance of a few important clocks from recent history. For many centuries, the daily rotation of Earth on its axis seemed to offer a uniform time base, but as time standards and measurement techniques improved, the length of a day was found to fluctuate and generally grow longer (which was attributed in part to tidal friction). Astronomers seeking a more stable unit of time chose the period of the orbital

Time & Frequency Division, National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305, USA.

*To whom correspondence should be addressed. E-mail: sdiddams@boulder.nist.gov

motion of Earth about the Sun (nominally 1 year) as the basis for the definition of the second. In 1956, the Ephemeris Second (1/31,556,925.9747 of the tropical year 1900) was formally adopted by the General Conference of Weights and Measures as the best measure of time.

Although the orbital motion of Earth in the solar system might be more uniform than the solar day, its period was impractically long for most purposes and was likely to suffer unpredictable changes and aging effects (hence, the definition of Ephemeris Time based on a particular solar year). Already when this definition of the second was adopted, scientists were investigating resonances or transitions in microscopic atomic systems as a more suitable means for defining time intervals and frequency. Many transitions between energy states in well-isolated atomic systems are highly immune to perturbations that would change the atomic resonance frequency ν_o, making these systems ideal candidates for clocks. Quantum mechanics dictates that the energies of a bound system (e.g., an electron bound to an atom) have discrete values. Hence, an atom or molecule can make a transition between two energy levels (E_1 and E_2) by the absorption or emission of energy in the form of electromagnetic radiation having the precise frequency $\nu_o = |E_1 - E_2|/h$, where h is the Planck constant. On the basis of this principle, most atomic frequency standards (atomic clocks) work by steering the frequency of an external oscillator to match a particular value of ν_o.

The first atomic clocks owe their genesis to the explosion of advances in quantum mechanics and microwave electronics before and during the Second World War. Much of the seminal work specific to clock development was done by Rabi. Although he may have suggested using cesium as the reference for an atomic clock as early as 1945, it was the inversion transition in the ammonia molecule at ~23.8 GHz that served as the reference for the first "atomic" clock in 1949 (*7*). In 1955, the first operational cesium atomic clock was built at the National Physical Laboratory, Teddington, UK (*8*). It was immediately noted that observations of the Moon over a period of several years would be required to determine Ephemeris Time with the same precision as was achieved in a matter of minutes by the first cesium clock (*9*). Although the fate of astronomically defined time seemed certain, more than a decade passed before the definition of the SI second was changed to be 9,192,631,770 cycles of the ground-state hyperfine splitting of the unperturbed cesium atom (*10*).

Cesium Clocks

Clocks are often characterized by their stability and accuracy. Stability is a measure of the degree to which the interval between "clock ticks" remains constant. Accuracy is a measure of how well the time between the clock ticks matches the defined second on the cesium hyperfine splitting. An accurate clock is necessarily stable over long intervals, but not all stable clocks are accurate.

On paper, clocks based on atomic processes are ideal, but there are fundamental as well as practical limitations to both their stability and accuracy. Atoms absorb or emit energy over a small range surrounding ν_o, not at ν_o alone. All other parameters being equal, the stability of an atomic clock is proportional to ν_o and inversely proportional to the small spread $\Delta\nu$ (linewidth) of absorption frequencies. This is more typically expressed in terms of the fractional frequency instability,

$$\sigma \propto \frac{\Delta\nu}{\nu_o}\frac{1}{S/N} \qquad (1)$$

where S/N is the signal-to-noise ratio, and high stability is equivalent to a smaller value of σ. From this expression we see that atomic clocks will generally benefit from operating at higher frequencies with transitions having narrow linewidths. In addition, Eq. 1 shows that the instability decreases with an increase in the S/N ratio with which the absorption signal is measured.

In the operation of an atomic clock, the atom must be illuminated by the electromagnetic radiation emitted from an external oscillator. Cesium clocks require a microwave oscillator, whereas the optical clocks discussed below require an optical oscillator (laser). The difficult challenge is to then tune the oscillator frequency to exactly match ν_o. There is always some ambiguity in this process because, as mentioned above, the resonance frequency has a nonzero linewidth associated with it. One factor that can limit the minimum observed linewidth of the reference transition is the time that the atom is in the radiation field. In those situations,

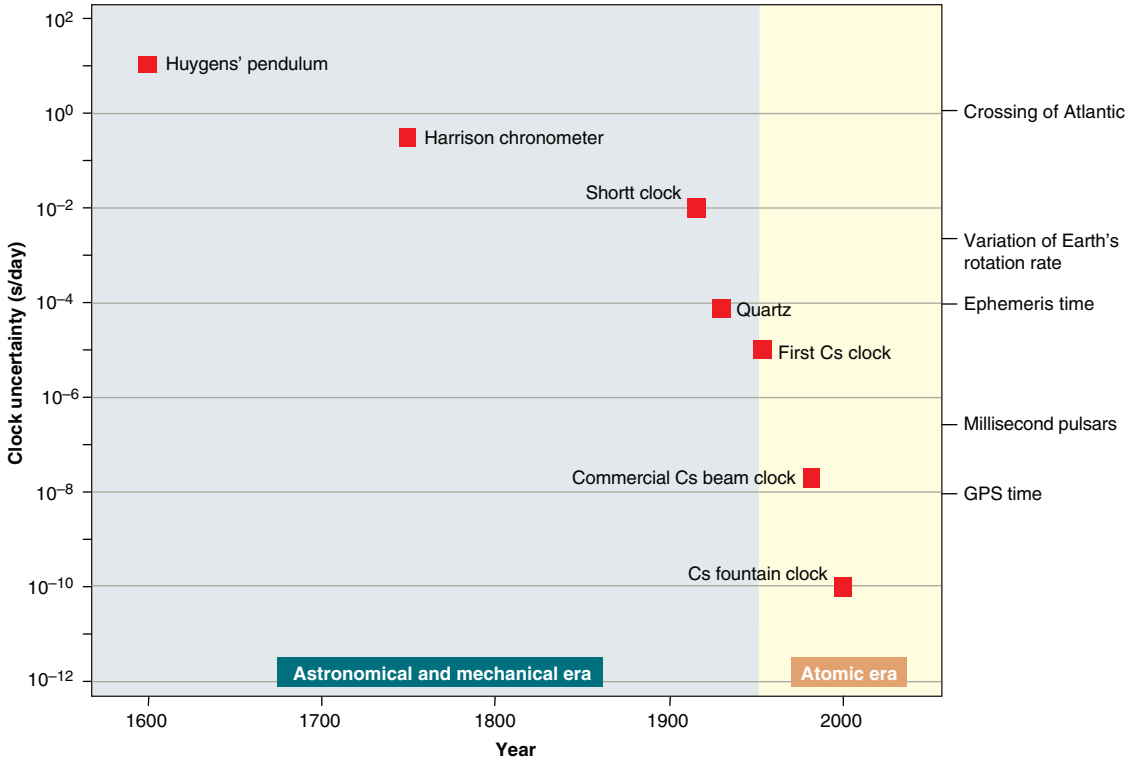


Fig. 1. Some of the major milestones in the improvement of clocks over the past 400 years. A clock with ~1 s uncertainty per day was required for crossing the Atlantic in 1750, and the Harrison chronometer demonstrated advancements that allowed such accurate navigation. The Shortt clock is the most accurate mechanical clock. Ephemeris time, as determined from astronomical observations, was accurate to ~0.1 ms per day. A clock based on a pulsar could have less than 1 μs of uncertainty per day. GPS time represents what typically can be achieved with a single receiver.

the observed linewidth of the resonance decreases as the measuring time increases.

Many other effects can act to degrade the stability and the accuracy of an atomic clock. The motion of the atoms introduces uncertainty by causing apparent shifts in the resonance frequencies (the Doppler effect). Similarly, collisions between atoms are a source of frequency shifts and linewidth broadening. Defects in the electronic measuring equipment as well as stray electromagnetic fields (including the ever-present thermal radiation) perturb the resonance frequency and introduce potential errors. Therefore, a good atomic clock must not only establish a steady periodic signal but must also minimize these potential errors.

A schematic of a conventional magnetically state-selected atomic beam cesium standard is shown in Fig. 2A. The design can be directly traced back to Rabi's and Ramsey's seminal works; essentially all commercial cesium atomic clocks use this general design, as do the cesium clocks in the GPS satellites. Moreover, this design provided the world with all of its primary frequency standards up until about 1990, and even today roughly 300 such clocks at more than 50 NMIs are included in the averaged international time scale known as Coordinated Universal Time (UTC) (*12*). The CS-1 clock of the Physikalisch-Technische Bundesanstalt (PTB) has the lowest stated systematic frequency inaccuracy of any clock of this type ever built, with

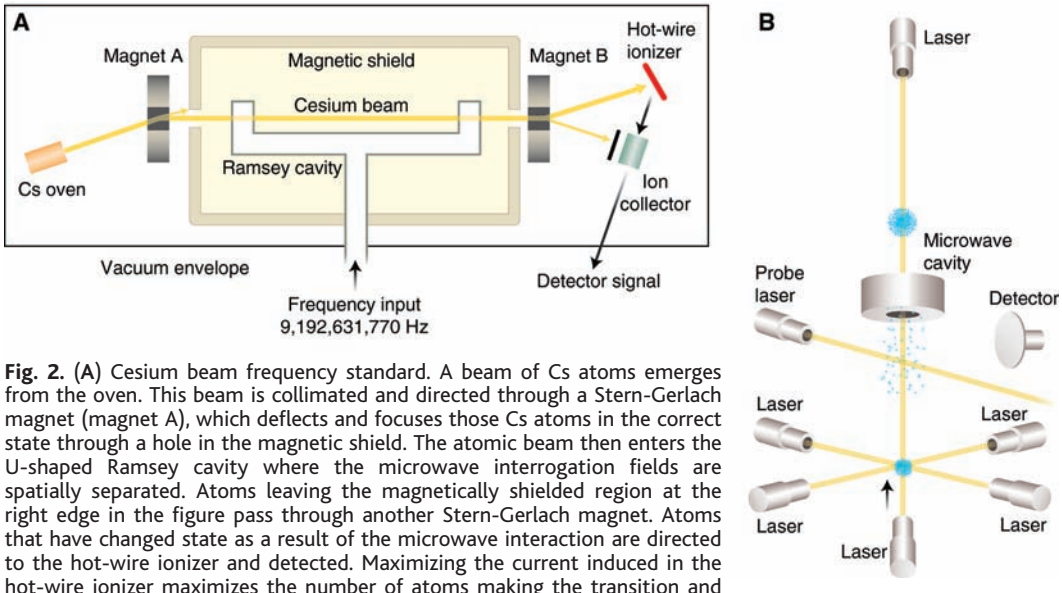


Fig. 2. (**A**) Cesium beam frequency standard. A beam of Cs atoms emerges from the oven. This beam is collimated and directed through a Stern-Gerlach magnet (magnet A), which deflects and focuses those Cs atoms in the correct state through a hole in the magnetic shield. The atomic beam then enters the U-shaped Ramsey cavity where the microwave interrogation fields are spatially separated. Atoms leaving the magnetically shielded region at the right edge in the figure pass through another Stern-Gerlach magnet. Atoms that have changed state as a result of the microwave interaction are directed to the hot-wire ionizer and detected. Maximizing the current induced in the hot-wire ionizer maximizes the number of atoms making the transition and thus assures that the frequency of the microwaves matches the atomic resonance frequency. (**B**) Cs fountain clock. The basic operation of the Cs fountain proceeds in a sequence of steps. First, a sample of $\sim 10^8$ cesium atoms is laser-cooled at the intersection of six laser beams to below 1 μK. These atoms are next "launched" upward at ~4 m/s by frequency detuning of the laser beams. The lasers are then turned off and the Cs atoms continue along their ballistic flight path. These atoms now enter the microwave cavity. The passage through the cavity on the way up provides the first pulse of the two-pulse (Ramsey) microwave interrogation sequence. The atoms reach apogee above the microwave cavity and eventually fall through the microwave cavity a second time. Atoms that have made a transition in state due to the interaction with the microwave field are detected optically with a laser.

as originally conceived by Zacharias in the 1950s (*14*). The idea was simple—to build a cesium beam clock vertically with one Ramsey interaction zone. Slow atoms in the cesium beam would traverse the microwave interaction zone traveling upward, reverse their velocity under the influence of gravity, and traverse the microwave interaction zone a second time traveling downward, resulting in Ramsey's two-pulse interaction scheme. With a ballistic flight traveling only a meter upward, the interaction time approaches 1 s instead of the 10 ms typical of beam clocks. Unfortunately, the scheme could not be realized; collisions between fast and slow cesium atoms in the beam very efficiently removed all the slow atoms Zacharias was counting on for the signal.

This idea was resurrected in the late 1980s when Chu and co-workers made the world's first working atomic fountain (*15*), which used laser cooling (*16*) to produce atoms at microkelvin temperatures. Researchers at the BNM-SYRTE (Bureau National de Metrologie–Systèmes de Référence Temps Espace) later built the first cesium primary frequency standard based on the fountain concept (*17*). Many other researchers in metrology laboratories around the world have built (or are building) laser-cooled cesium fountain primary frequency standards similar to the schematic design shown in Fig. 2B. The resulting interaction time, ~1 s, allows fountain-based frequency standards to achieve much lower inaccuracy than beam standards. The present systematic inaccuracy of the NIST fountain clock (NIST-F1) is $\delta\nu/\nu_o < 4 \times 10^{-16}$, with other fountain frequency standards having similar or only slightly larger inaccuracy (*18–21*).

At this level of performance, one finds that the limitations arise from a variety of interesting physical interactions that are fundamental in nature. For example, given a certain number of cesium atoms, the instability of the clock output is limited by quantum-mechanical measurement statistics, whereas the accuracy of the clock is affected primarily by uncertainty in the ambient blackbody radiation, knowledge of the local gravitational potential, and collisions between the cesium atoms. In the case of NIST-F1, the cesium atoms are bathed in the blackbody radiation from the surrounding room at 300 K. This results in a correctable shift of

In the early work of Rabi, the atomic resonance was interrogated with the radiation from one long microwave pulse. This provided the needed long interaction time between the atom and microwave field, but led for various reasons to the output frequency of the standard being subject to Doppler shifts and other sensitivities. Ramsey's method of separated oscillatory fields (*11*) provided a critical improvement that has been adopted by all modern primary frequency standards. In Ramsey's method, the microwave excitation is done in two relatively short pulses at the beginning and end of the interaction zone. This two-pulse process (now known as Ramsey interrogation) reduces these sensitivities by factors of 10 to 100 or more.

$\delta\nu/\nu_o \sim 8 \times 10^{-15}$ (*13*). A slight twist in the design of a cesium beam standard is to replace the state-selection magnets (magnets A and B in Fig. 2A) with lasers that optically pump the cesium atoms into specific energy states, thereby making state selection and detection more efficient. This provides some improvement, and the best optically pumped cesium beam standards have frequency inaccuracies on the order of $\delta\nu/\nu_o \approx 3 \times 10^{-15}$. This is only slightly more accurate than the best magnetically selected thermal cesium beams, because both versions are limited fundamentally by the large atom velocities and the resulting short interaction time.

A solution to this problem is to use slowly moving cesium atoms in a fountain geometry,

$\delta\nu/\nu_o = 2 \times 10^{-14}$, but the uncertainty in the correction is $\delta(\delta\nu/\nu_o) \approx 2.5 \times 10^{-16}$, corresponding to an uncertainty of 1 K in the temperature of the thermal radiation. At a similar level, the gravitational redshift moves the clock frequency by approximately $\delta\nu/\nu_o = 10^{-16}$ per meter change in the elevation of the cesium with respect to sea level. Furthermore, in NIST-F1, collisions between cesium atoms cause a frequency shift with an uncertainty of $\delta\nu/\nu_o \approx 1.5 \times 10^{-16}$. The next generation of cesium fountain clocks are being designed to overcome some of these limitations, with the intention of reducing the total uncertainty in accuracy to the level of $\delta\nu/\nu_o \approx 10^{-16}$. After that, the best cesium clocks might have to operate in orbit around Earth, where the weightless environment would allow interaction times of 5 to 10 s with slow cesium atoms, resulting in projected fractional uncertainties as low as 5×10^{-17} (*22*, *23*).

Optical Frequency Standards and Clocks

Although present-day cesium microwave frequency standards perform at an already remarkable level, a new approach to timekeeping based on optical atomic transitions promises still greater improvements. By using optical ($\nu_0 \sim 10^{15}$ Hz) rather than microwave ($\nu_0 \sim 10^{10}$ Hz) frequencies, such a clock operates with a much smaller unit of time (comparable to using a second rather than a day as the basic unit). One can see from Eq. 1 that basing a standard on a transition in the optical rather than the microwave region of the spectrum could in principle lead to an enormous reduction in instability. Optical standards should be considerably more accurate as well, as several key frequency shifts are fractionally much smaller in the optical domain. Moreover, the investigation of these shifts will be greatly accelerated by the much smaller instability of the optical standards.

These potential advantages were recognized in the early days of frequency standards, but a coherent light source was needed to serve as the local oscillator. Soon after the first demonstrations of lasers in the early 1960s, the search began for relevant transitions on which to stabilize the laser frequencies. Some of the earliest stabilized lasers were based on transitions in the well-known helium-neon laser, whose frequency was locked to absorption lines in methane (in the infrared) or iodine (at 633 nm) molecules (*24*, *25*). With the advent of tunable lasers, frequency standards based on molecules, neutral atoms, and ions now exist throughout the visible spectrum. These standards extend to the near-infrared wavelengths as well, including important references for the densely spaced fiber communications frequency channels in the 1.3- and 1.5-μm region (*26*).

However, optical standards did not truly begin to realize the potential gains expressed in Eq. 1 until the past decade, when several critical technologies reached maturity. First, advances in laser cooling techniques of atoms and ions made it possible to cool a variety of atoms and ions (including those with narrow clock transitions) to millikelvin temperatures and below (*16*). The use of laser-cooled atomic samples enables the extended interaction times (along with reduced Doppler shifts) required to observe a narrow transition linewidth ($\Delta\nu$ in Eq. 1). Second, to resolve narrow linewidths, probe lasers need to be spectrally pure. Recent improvements in laser stabilization based on environmentally isolated optical reference cavities have enabled laser linewidths at the subhertz level to be achieved (*27*, *28*). Finally, and perhaps most critically, a simple means for counting optical frequencies and linking them to other frequencies in the microwave and optical domains became available with the development of the mode-locked femtosecond (10^{-15} s) laser frequency comb (*29*, *30*). As a result of these improvements, there is renewed enthusiasm in this field, with many groups now racing to develop new standards and clocks based on various transitions.

When developing a new standard, the choice of the atomic reference transition itself is clearly one of the most critical aspects. For state-of-the-art performance, one desires a narrow transition that is extremely insensitive to external perturbations such as frequency shifts caused by magnetic and electric fields. Such transitions (often called clock transitions) exist in both ions and neutral atoms, and the choice currently comes down to working with a few trapped ions or a large number of neutral atoms.

A trapped ion can be laser-cooled to the zero point of its motion, thereby suppressing Doppler effects that can shift the resonance frequency (*31*). Moreover, an ion can be probed while trapped, hence long interaction times can be achieved. Although these factors give trapped-ion standards excellent prospects for high accuracy, their S/N ratio is limited, because in most cases ion-ion interactions become a limiting factor if more than a few ions are in the trap. Nonetheless, outstanding performance has been demonstrated with optical transitions in a variety of single-ion systems, including Hg^+, Yb^+, Sr^+, and In^+ (*32*–*34*).

Large numbers of neutral atoms can be probed simultaneously, which enables an extremely high S/N ratio and potentially very high stability. Unfortunately, methods of confining neutral atoms, although efficient for collecting atomic samples, can lead to shifts for the reference transition. Thus, the atoms typically need to be released during the probe cycle, which leads to troublesome Doppler-related systematic shifts and limited interaction times. Still, excellent performance has been demonstrated with several laser-cooled neutral-atom standards (*35*). Moreover, a possible solution to the confinement problem has been proposed, which

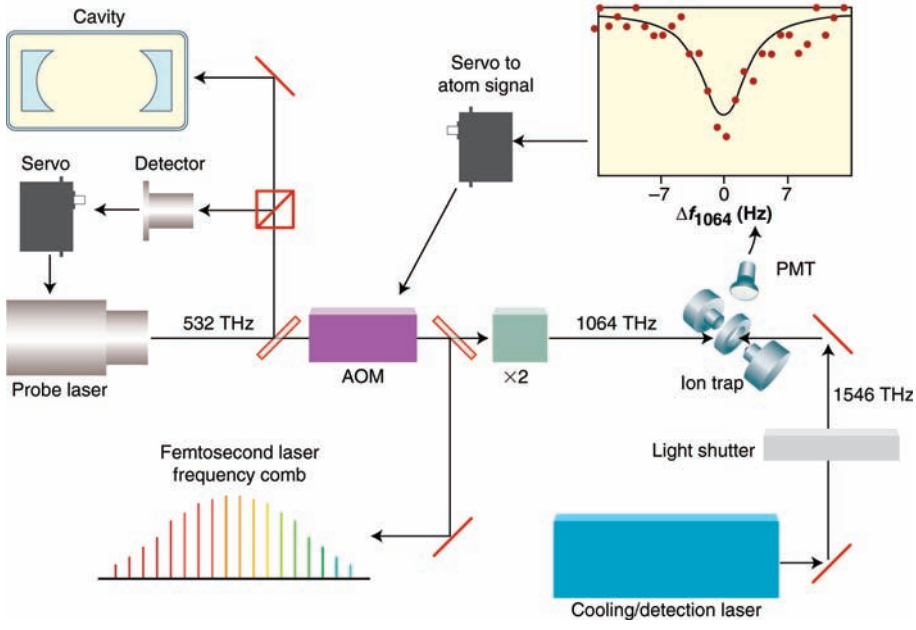


Fig. 3. Main components of an optical atomic clock. The probe laser, whose frequency is prestabilized on an optical cavity, is used to excite transitions in the laser-cooled trapped ion. A servo system uses the signal from the ion to keep the probe laser frequency centered on resonance. Light is sent to the femtosecond-laser frequency comb, which enables counting of the clock cycles.

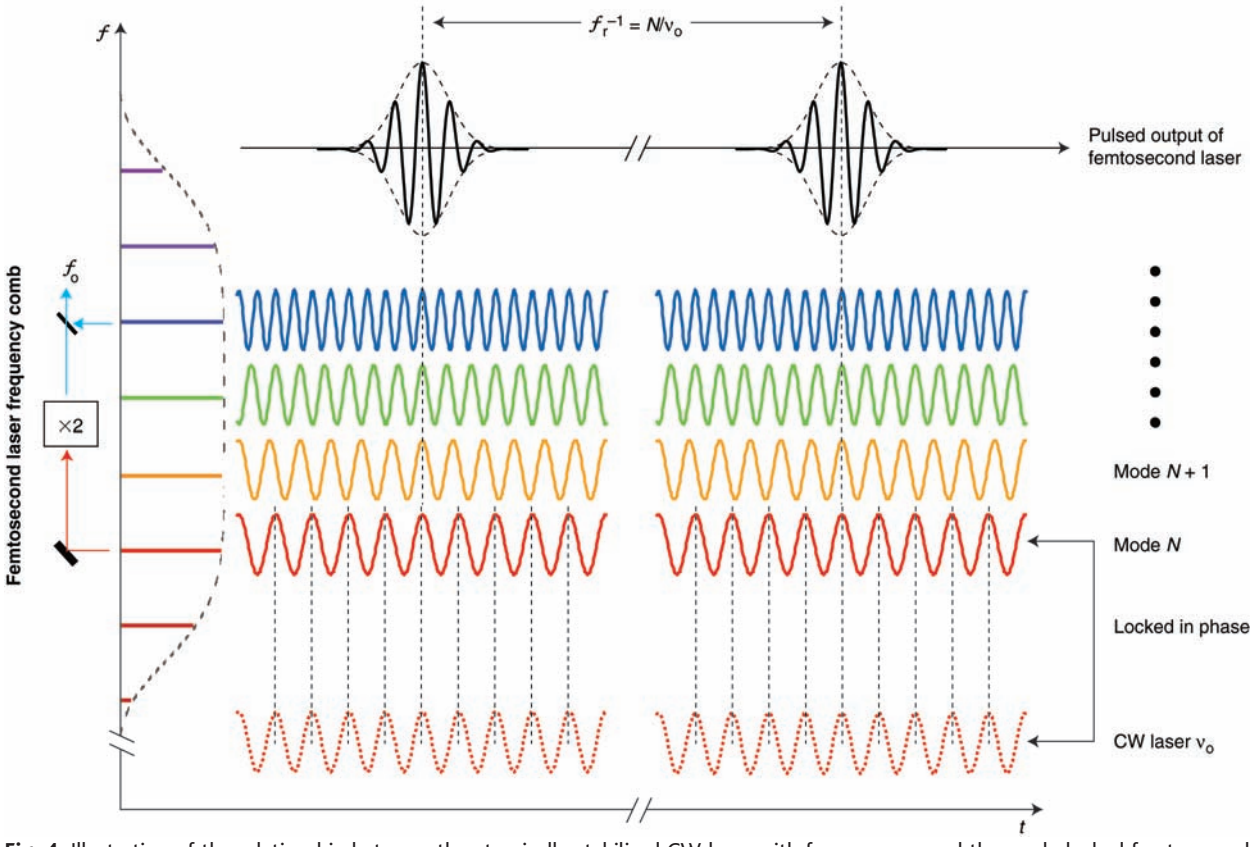


Fig. 4. Illustration of the relationship between the atomically stabilized CW laser with frequency ν_o and the mode-locked femtosecond laser pulses. When the femtosecond laser comb is phase-locked to ν_o, the frequency of the pulse repetition rate is simply a rational fraction of ν_o. For clarity, only a few of the femtosecond laser comb elements are shown here. Actual devices might have ~500,000 modes separated by 1 GHz.

involves loading neutral atoms into a specially designed laser lattice (*36*). The wavelength of the lattice laser beams can be chosen so as to minimize shifts of the clock transition. In this way, one could have the long interaction times and small Doppler shifts associated with ions, along with the good S/N ratio achievable with large numbers of neutral atoms. Laser-cooled neutral-atom systems currently under development for optical clocks include Ca, Sr, Yb, Mg, and H (*32*–*34*).

The single-ion Hg^+ system, a typical state-of-the-art optical frequency standard, is shown in Fig. 3. A measurement cycle commences with laser cooling of the ion, followed by a probe period during which the ion is excited by a pulse from the probe laser, whose frequency is locked to a narrow resonance line of a high-finesse Fabry-Perot cavity with an electronic feedback loop. An optical cavity can serve as near-ideal short-term reference, providing a comb of resonance lines that can be narrow and resolved with high S/N ratio. As the frequency of a given cavity line depends on the spacing of the mirrors, it is essential to isolate the cavity spacer that determines the mirror separation from environmental perturbations. Thus, the cavity spacer is usually made from a special material and placed in an isolation chamber. Using this approach, lasers with linewidths of 1 Hz or less can be maintained for durations of tens of seconds (*28*). A frequency-shifting device such as an acousto-optic modulator (AOM) then adjusts the laser frequency so that it is near the atomic resonance. Near-resonant probe pulses induce excitation in the atomic sample, which is detected by collecting atomic fluorescence on a photomultiplier tube. So-called ''shelving detection'' schemes based on a strong transition (usually the laser cooling transition) are usually used to enhance the detected signals, thereby enabling atom shot noise–limited performance (*37*, *38*).

Because we are working with a quantum system, a single measurement on a single atom cannot tell us the frequency of the probe (''clock'') laser, but only whether the atom was excited by the probe pulse. For a single ion, it is necessary to average the excitation over many probe cycles to determine the excitation probability, which in turn can be related to the frequency of the laser. Alternatively, one can use a single probe cycle to excite a large number of atoms, effectively performing the averaging simultaneously (as one typically does with neutral atom clock samples). As an example, we show in Fig. 3 an averaged excitation spectrum obtained by scanning the probe laser frequency over the Hg^+ clock resonance near 282 nm (this signal has a line $Q = \nu_o/\Delta\nu = 10^{14}$, the highest demonstrated in the microwave or optical region of the spectrum) (*39*). With appropriate modulation techniques, a spectroscopic signal suitable for locking the laser to the center of this resonance can be generated. This second stage of locking uses feedback to the AOM frequency to keep the laser frequency fixed on the atomic resonance, thereby suppressing residual cavity drifts. With such a lock, the probe laser frequency can now exhibit excellent long-term performance and be used as an optical frequency standard.

Several such standards have been constructed using either single ions or about a million neutral atoms, many of which now achieve inaccuracies of 10^{-14} or better, including Hg^+, Yb^+, Sr^+, Ca, and H. It is anticipated that in the next year a few optical standards will be evaluated at the 10^{-15} level and below. It has been predicted that the systematic effects in single-ion optical frequency standards could be controlled at a level that would permit uncertainties approaching 10^{-18} (*38*). Reaching such a level will necessarily require short-term instability

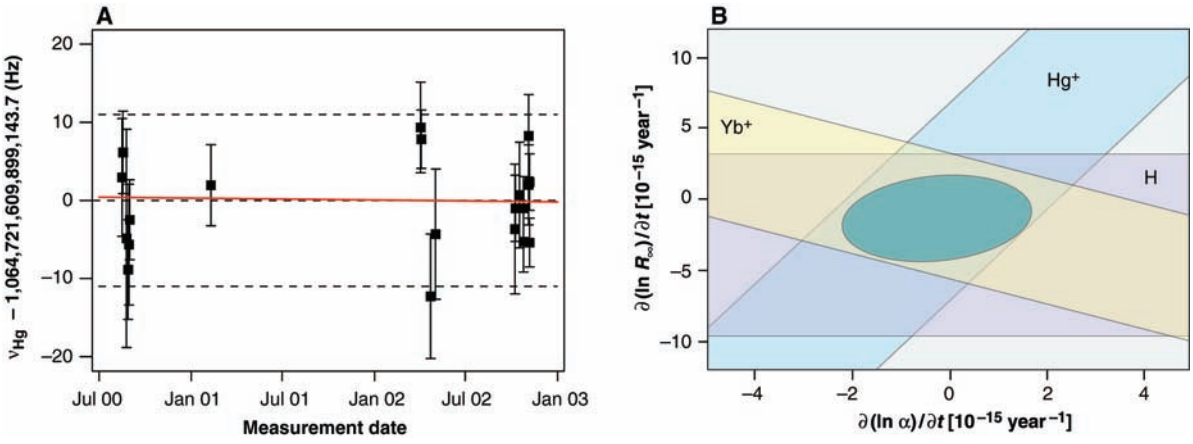


Fig. 5. (**A**) Measurement of the Hg^+ optical clock transition with respect to the Cs ground state hyperfine splitting that defines the SI second. The plot shows the deviation of each measurement from the weighted average value with its statistical error bar and a linear fit to these data (red line). The total systematic uncertainty is represented by the dashed line at ±11 Hz. Using this measurement in combination with other recent comparisons of H and Yb^+ to the Cs ground state hyperfine splitting allows one to constrain possible variations of α and R_∞ as shown in (**B**). The central ellipse represents the region consistent with the combination of all three experiments [see (*56*) for details].

at or below 10^{-15} for 1 s of signal averaging if inordinately long averaging times are to be avoided.

Femtosecond Laser Frequency Divider

To generate an interval of time with an optical standard (and thereby compare it to existing microwave atomic clocks), one requires a reliable and straightforward means of counting the extremely rapid optical oscillations that occur with a period of ~1 fs—a time scale much too fast for any conventional electronic counters. In the past, solutions to this counting problem involved complicated approaches based on harmonic frequency chains (*40*) or interval bisection (*41*). However, the scene changed in 1999 when key experiments by the Hänsch group demonstrated that the frequency comb associated with a femtosecond mode-locked laser could be used for counting optical frequencies (*29*). Indeed, in just a matter of years the new femtosecond comb technology has fully replaced other laboratory technologies that existed for decades.

The basic function of the femtosecond laser in an optical clock is to provide a phase-coherent link between the uncountable optical reference frequency and the more accessible microwave domain. This operation can be understood by considering the underlying frequency comb structure associated with the femtosecond laser (Fig. 4). A femtosecond laser produces short pulses from a superposition of phase-locked cavity modes. In the frequency domain, these modes form a comb of evenly spaced oscillators. The spacing of the modes is given by the repetition rate f_r at which pulses are emitted from the mode-locked laser. Moreover, a difference in group and phase delays inside the femtosecond laser cavity leads to a frequency offset, f_o, of the comb elements from being exact harmonics of f_r. The relationship between these two microwave frequencies (f_o, f_r) and the nth element of the optical frequency comb is given by $\nu_n = nf_r + f_o$, where n is an integer.

In general, the group and phase delays in the femtosecond laser fluctuate and are not calculable at a level that permits a high-precision determination of f_o. Therefore, f_o must be directly measured, and an elegant means of accomplishing this involves using nonlinear frequency generation to compare different regions of the frequency comb (*42*). For example, if the laser spectrum covers more than one octave, then the comb elements at the low-frequency end of the spectrum can be doubled in a nonlinear crystal and subsequently heterodyned against the high-frequency components of the comb to yield f_o. An important advance in this respect is the generation of octave-spanning spectra with low-power Ti:sapphire lasers in microstructured fibers (*43*, *44*), or by direct generation from the laser itself (*45*–*47*). Once measured, f_o can then be locked at a fixed frequency with the use of servo-control techniques. If, for example, it is set to zero, then each comb element is an exact harmonic of f_r. Phase-locking one element ($n = N$) of the frequency comb to the low-noise continuous-wave (CW) laser that has itself been steered to the atomic resonance leads to the desired result for optical to microwave conversion: $f_r = \nu_o/N$ (Fig. 4). With such a scheme, the complete connection from the optical to microwave domain has now been demonstrated—starting with either a single ion or a few million neutral atoms and ending with the 1 GHz optical pulse train from a femtosecond laser (*48*). The excess noise of this division process has been verified at a level sufficiently low to support the best current optical frequency standards (*49*, *50*).

Outlook

Today's cesium atomic clocks have timing uncertainty at the level of ~35 ps per day, and optical atomic clocks might someday have uncertainties near 100 fs per day. When most people would get by just fine with a clock that is accurate to a few seconds per day, it is worth asking why scientists, or society for that matter, might want still better clocks. In the 1950s, when microwave atomic clocks based on cesium were first developed, the situation was in some ways similar to where we find ourselves today. Those first atomic clocks were quickly recognized as a great improvement over the existing clocks, yet at the same time they were mainly a tool of scientific interest. At that point, few people would imagine that just 40 years into the future a constellation of satellites containing cesium (and rubidium) atomic clocks would circumnavigate the globe and provide accurate time and position to all people below. The GPS and its constituent atomic clocks are now an integral part of our lives. Similarly, today's voice and data communications systems that are synchronized with atomic clocks would likely have been viewed as science fiction in 1955. In the next 50 years, there is every reason to expect that improved microwave clocks and optical clocks will find numerous applications in communications and navigation—the two areas that have throughout history advanced in parallel with improving time standards. The very stable optical clock ticks may be especially useful for tracking and communication between satellites and spacecraft in the much greater expanses beyond our planet. For example, an optical clock with femtosecond instability would provide potential ranging uncertainty at the micrometer level over millions of kilometers.

For the immediate future, it is already clear that the most advanced clocks will provide interesting new scientific avenues to study our universe, pushing the limits on tests of the most fundamental physical laws to new levels. This includes tests of general relativity and searches for violations of the isotropy of space or a preferred reference frame. Fundamental symmetries between matter and antimatter could be investigated through the comparison of optical clocks, as has been proposed for the 1*s*-2*s* transitions in both hydrogen and anti-hydrogen (*51*). To date, optical and microwave frequency standards have already been used in some of the most accurate determinations of the fine structure constant α and the Rydberg constant R_∞ [e.g., (*52*, *53*)], and laboratory comparisons of clocks based on different atomic transitions are now providing some of the most stringent constraints of the possible variation of fundamental constants (*54*, *55*). An example of this kind of experiment is given in Fig. 5A, which shows the measurement of the Hg^+ optical clock transition at ~1064 THz (282 nm) in terms of the cesium hyperfine splitting as realized by NIST-F1 (*54*). Over a period of ~3 years, there is no measurable divergence in the ratio of the output frequencies of these two clocks, constraining a fractional variation of $g_{Cs}(m_e/m_p)\alpha^6$ to be less than 7×10^{-15} per year, where m_e/m_p is the electron-to-proton mass ratio and g_{Cs} is the ^{133}Cs nuclear *g* factor. Assuming any variation comes only from the α^6 factor, the data constrain any possible linear fractional variation of α to be less than 1.2×10^{-15} per year. The combination of these data with other recent clock comparison experiments has resulted in similar constraints being placed on other fundamental constants, as summarized in Fig. 5B (*56*, *57*).

It seems clear that future atomic clocks will continue to subdivide the second into still smaller units of time. But in contrast to Mermin's original concern, it is more likely that the femtosecond, attosecond (10^{-18} s), or zeptosecond (10^{-21} s) will be considered utterly alien impositions on our macroscopic world, while nonetheless proving to be useful bookkeeping units in the continued quest to better understand the inner workings of the microscopic world.

References and Notes

1. N. D. Mermin, *Phys. Today* **54**, 11 (2001).
2. J. H. Taylor, in *Nobel Lectures, Physics 1991–1995*, G. Ekspong, Ed. (World Scientific, Singapore, 1997), pp. 73–91.
3. R. F. C. Vessot *et al.*, *Phys. Rev. Lett.* **45**, 2081 (1980).
4. R. E. Beehler, *Proc. IEEE* **55**, 792 (1967).
5. N. F. Ramsey, *J. Res. Nat. Bur. Stand.* **88**, 301 (1983).
6. C. Audion, B. Guinot, *The Measurement of Time: Time, Frequency and the Atomic Clock* (Cambridge Univ. Press, Cambridge, 2001).
7. H. Lyons, *NBS Tech. News Bull.* **33**, 17 (1949).
8. L. Essen, J. V. L. Parry, *Nature* **176**, 280 (1955).
9. E. C. Bullard, *Nature* **176**, 282 (1955).
10. Resolution 1, 13e Conférence Générale des Poids et Mesures, *Metrologia* **4**, 41 (1968).
11. N. F. Ramsey, *Phys. Today* **33**, 25 (1980).
12. T. J. Quinn, *Proc. IEEE* **79**, 894 (1991).
13. A. Bauch, B. Fischer, T. Heindorff, R. Schröder, *Metrologia* **35**, 829 (1998).
14. N. F. Ramsey, *Molecular Beams* (Clarendon, Oxford, 1956).
15. M. Kasevich, E. Riis, S. Chu, R. DeVoe, *Phys. Rev. Lett.* **63**, 612 (1989).
16. H. J. Metcalf, P. van der Straten, *Laser-Cooling and Trapping* (Springer, New York, 1999).
17. A. Clairon *et al.*, in *Proceedings of the 5th Symposium on Frequency Standards and Metrology,* J .C. Berguist, Ed. (World Scientific, London, 1996), pp. 49–59.
18. T. P. Heavner, S. R. Jefferts, E. A. Donley, J. H. Shirley, T. E. Parker, *IEEE Trans. Instrum. Meas.*, in press.
19. S. Weyers, U. Hübner, R. Schröder, Chr. Tamm, A. Bauch, *Metrologia* **38**, 343 (2001).
20. S. Bize *et al.*, in (*32*), pp. 53–63.
21. F. Levi, L. Lorini, D. Calonico, A. Godone, *IEEE Trans. Ultrason. Ferroelect. Freq. Control* **51**, 1216 (2004).
22. P. Laurent *et al.*, in (*32*), pp. 241–252.
23. T. P. Heavner *et al.*, in (*32*), pp. 253–260.
24. J. L. Hall, C. J. Borde, K. Uehara, *Phys. Rev. Lett.* **37**, 1339 (1976).
25. J.-M. Chartier, A. Chartier, *Proc. SPIE* **4269**, 123 (2001).
26. S. L. Gilbert, W. C. Swann, T. Dennis, *Proc. SPIE* **4269**, 184 (2001).
27. J. L. Hall, *Science* **202**, 147 (1978).
28. B. C. Young, F. C. Cruz, W. M. Itano, J. C. Bergquist, *Phys. Rev. Lett.* **82**, 3799 (1999).
29. Th. Udem *et al.*, *Phys. Rev. Lett.* **82**, 3568 (1999).
30. S. A. Diddams *et al.*, *Phys. Rev. Lett.* **84**, 5102 (2000).
31. F. Diedrich, J. C. Bergquist, W. M. Itano, D. J. Wineland, *Phys. Rev. Lett.* **62**, 403 (1989).
32. P. Gill, Ed., *Proceedings of the 6th Symposium on Frequency Standards and Metrology* (World Scientific, Singapore, 2002).
33. *Proceedings of the Conference on Precision Electromagnetic Measurements* (IEEE, Piscataway, NJ, 2004).
34. J. Ye, H. Schnatz, L. W. Hollberg, *IEEE J. Select. Top. Quantum Electron.* **9**, 1041 (2003).
35. U. Sterr *et al.*, *C. R. Phys.*, in press.
36. H. Katori, M. Takamoto, V. G. Pal'chikov, V. D. Ovsiannikov, *Phys. Rev. Lett.* **91**, 173005 (2003).
37. D. J. Wineland, J. C. Bergquist, W. M. Itano, R. E. Drullinger, *Opt. Lett.* **5**, 245 (1980).
38. H. G. Dehmelt, *IEEE Trans. Instrum. Meas.* **31**, 83 (1982).
39. R. Rafac *et al.*, *Phys. Rev. Lett.* **85**, 2462 (2000).
40. D. A. Jennings, K. M. Evenson, D. J. E. Knight, *Proc. IEEE* **74**, 168 (1986).
41. H. R. Telle, D. Meschede, T. W. Hänsch, *Opt. Lett.* **15**, 532 (1990).
42. H. R. Telle *et al.*, *Appl. Phys. B* **69**, 327 (1999).
43. J. K. Ranka, R. S. Windeler, A. J. Stentz, *Opt. Lett.* **25**, 25 (2000).
44. D. J. Jones *et al.*, *Science* **288**, 635 (2000).
45. U. Morgner *et al.*, *Phys. Rev. Lett.* **86**, 5462 (2001).
46. A. Bartels, H. Kurz, *Opt. Lett.* **27**, 1839 (2002).
47. T. Fortier, D. J. Jones, S. T. Cundiff, *Opt. Lett.* **28**, 2198 (2003).
48. S. A. Diddams *et al.*, *Science* **293**, 825 (2001).
49. L.-S. Ma *et al.*, *Science* **303**, 1843 (2004).
50. A. Bartels *et al.*, in *OSA Trends in Optics and Photonics Series (TOPS) Vol. 96, Conference on Lasers and Electro-Optics (CLEO)* (Optical Society of America, Washington, DC, 2004), paper CPDC10.
51. J. Walz *et al.*, *Appl. Phys. B* **77**, 713 (2003).
52. F. Biraben *et al.*, in *The Hydrogen Atom: Precision Physics of Simple Atomic Systems*, S. V. Karshenboim, F. S. Pavone, F. Bassani, M. Inguscio, T. W. Hänsch, Eds. (Springer-Verlag, Berlin, 2001), pp. 17–41.
53. Th. Udem *et al.*, in *The Hydrogen Atom: Precision Physics of Simple Atomic Systems*, S. V. Karshenboim, F. S. Pavone, F. Bassani, M. Inguscio, T. W. Hänsch, Eds. (Springer-Verlag, Berlin, 2001), pp. 125–144.
54. S. Bize *et al.*, *Phys. Rev. Lett.* **90**, 150802 (2003).
55. H. Marion *et al.*, *Phys. Rev. Lett.* **90**, 150801 (2003).
56. E. Peik *et al.*, *Phys. Rev. Lett.* **93**, 170801 (2004).
57. M. Fischer *et al.*, *Phys. Rev. Lett.* **92**, 230802 (2004).
58. We thank L. Hollberg, D. Wineland, and M. Lombardi for their assistance and thoughtful comments, and E. Peik and his colleagues at the PTB for providing Fig. 5B.

REVIEW

The Route to Atomic and Quantum Standards

Jeff Flowers

Over the past half-century, there has been a shift away from standards based on particular artifacts toward those based on physical effects, the most stable being based on quantum properties of systems. This change was proposed at the end of the 19th century but is still not complete at the start of the 21st. We discuss how this vision has been implemented through recent advances in science and metrology and how these may soon lead to an SI system finally free from artifact standards, with a consistency based on fundamental constants.

Quantities, Units, and Standards

To investigate any physical phenomena, we must make measurements, communicate them to others, and record them in a way that will be understandable in the future. To do so, a system of quantities and units is required. Measurement is a comparison process in which the value of a quantity is expressed as the product of a value and a unit; that is,

$$\text{Quantity} = \{\text{numerical value}\} \times [\text{unit}] \quad (1)$$

where the unit is an agreed-upon value of a quantity of the same type. The concept of a quantity such as length is independent of the associated unit; the length is the same whether it is measured in feet or meters. A standard is a

National Physical Laboratory, Teddington, Middlesex TW11 0LW, UK. E-mail: jeff.flowers@npl.co.uk

physical realization of the definition, with an agreed-upon value to be used as a reference.

Systems of measurement have been developed since ancient times. During the 19th and 20th centuries, the development of global trade in goods and international communication between scientists led to the development of an internationally agreed-upon system of units based on standards endorsed by international treaty. This has virtually replaced, especially in scientific and technical use, the confusing multiplicity of units and standards that existed beforehand and were often only in local use (*1*). Over time, as understanding and technology developed, this system was enhanced and developed into the Système International d'Unites (SI), formalized in 1960, that is now in global use. The history leading to the establishment of the SI and changes made by decisions of the Conférence Générale des Poids et Mesures (CGPM), the decision-making body, are recorded in the SI brochure, now in its 7th edition (*2*, *3*).

Early Proposals for Units Based on Physical Phenomena

Early ''natural'' unit systems used Earth's size and rotation rate and the density of water as units, but these are difficult to realize, so for practical purposes the international standards of mass and length were originally artifacts preserved in vaults. Later, modern artifacts, standard cells, and standard resistors, for example, were invented. It was recognized early on (as in the quotation of Maxwell reproduced below) that a natural system of standards could be based on atomic phenomena that were reproducible and constant and not subject to the vagaries of man and the environment.

> *The earth has been measured as a basis for a permanent standard of length, and every property of metals has been investigated to guard against any alteration of the material standards when made. To weigh or measure any thing with modern accuracy, requires a course of experiment and calculation in which almost every branch of physics and mathematics is brought into requisition. Yet, after all, the dimensions of our earth and its time of rotation, though, relatively to our present means of comparison, very permanent, are not so by any physical necessity. The earth might contract by cooling, or it might be enlarged by a layer of meteorites falling on it, or its rate of revolution might slowly slacken, and yet it would continue to be as much a planet as before. But a molecule, say of hydrogen, if either its mass or its time of vibration were to be altered in the least, would no longer be a molecule of hydrogen. If, then, we wish to obtain standards of length, time, and mass which shall be absolutely permanent, we must seek them not in the dimensions, or the motion, or the mass of our planet, but in the wave-length, the period of vibration, and the absolute mass of these imperishable and unalterable and perfectly similar molecules.*
>
> —James Clerk Maxwell, 1890 (*4*), p. 225

SI is a practical system, based on atomic or quantum standards where appropriate, but based on artifacts where utility dictates. It cannot be led by philosophy beyond the bounds of present-day science and technology. The change from artifact to atomic or quantum standards has been a guiding principle, but improved realization and stability are prerequisites of advances.

Role of the Fundamental Constants in Unit Systems

It is common in theoretical work to use the fundamental constants of physics as units. This is often seen in the form of an expression such as ''let $\hbar = e = c = 1$'' or ''let $e = m_e = \hbar = 1$'' at the start of a paper. These expressions define an alternative unit system with different base quantities from those of the SI. They could be written as $\{e\} = 1$, and so on, using the notation of Eq. 1, to show the meaning; for example, $e = 1$[e], the value of the electron charge e, is one e unit. Nowadays, the metrologist is aiming to do in practice what the theoretician is doing in principle, basing the system of units on fundamental constants or atomic systems. These are invariant both on a practical scale and as far as can be measured in the laboratory (*5*, *6*). The possible variability of the fundamental constants currently has a high profile, following evidence for variation in the fine-structure constant over cosmological time (*7*), although this is now disputed (*8*). Variations of the constants would imply new physics that would require changes in the basis of our unit system but would not impact practical measurement at current accuracies. For comparison of a theory with experiment, the conversion must be made to practical (preferably SI) units, and this requires values for the appropriate fundamental constants in SI units at the requisite accuracy.

Length, Time, and Frequency

Mass, length, and time are the quantities that have been measured since earliest times, and they were the basis of early unit systems. We will defer the discussion of mass until later in this article because of the present link with electrical metrology, which we discuss below.

Because of its unrivaled accuracy, frequency measurement is key in many aspects of modern metrology. Early clocks were mechanical, with a pendulum as a frequency standard; the practical change to an atomic system came in 1955 with the cesium clock developed by Essen (*9*) (Fig. 1A). By 1968, further refinement and testing led to the redefinition of the second from one based on the rotation of Earth to an atomic one: ''The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom'' [13th CGPM, 1967 (*10*)].

The time lag in adopting the atomic standard as the definition reflects the work that is necessary to give confidence in the superiority of a new method and in ensuring that the new value adopted is as close as possible to the old one.

Although the cesium frequency cannot presently be explicitly written in terms of fundamental constants because of the complexity of the atomic theory required, it is a quantum system that will have the stability associated with fundamental constants. The uncertainty in calculating this frequency is many orders of magnitude away from its measurement uncertainty. The Rydberg constant could be considered the natural fundamental constant–based unit of frequency. It is determined with a relative uncertainty of 6.6×10^{-12}, which is currently the limit at which an atomic frequency can be calculated from fundamental constants (*11*).

The choice of the cesium definition was a good one in the sense that the technology, although superior to the alternative clocks of the day, still had much room for improvement, and the definition has endured to this day, during which time its practical realization has improved by five orders of magnitude. However, the time is approaching when optical frequency standards will have accuracies and stabilities superior to the best microwave cesium standards. Then it will be necessary to revisit the definition of the second. There are a number of candidate optical frequency standards (*12*), but at present no particular standard is clearly superior to the others.

The development of modern frequency metrology has led to a measurement capability of astounding accuracy. As it is expanded on in another article in this issue (*12*), I will consider here in detail only one aspect: the change in status of length and the meter.

The speed-of-light definition of the meter. A clear example of the link between fundamental constants and the units is the adoption of the speed-of-light definition of the meter. The meter was originally defined as the length of a prototype meter bar intended to be 1/10,000,000 of the length of a quadrant of Earth. By 1960, the development of interferometry allowed an atomic redefinition of the meter in terms of the wavelength of

light from a specific source, the krypton lamp. With the invention of the laser, length measurement by interferometry was radically improved and the krypton standard was not accurate enough. The meter definition could then have been revised using the wavelength of a specified stabilized laser. However, the progress in understanding the metrological importance of the speed of light, along with the progress in its accurate measurements (Fig. 1B), led to the change from defining the meter in terms of the wavelength of light from a specific source, to a fundamental constant–based definition in which the speed of light is a defined quantity (*13*). Thus, the definition is now: ''The metre is the length of path travelled by light in vacuum during a time interval of 1/299 792 458 of a second'' [17th CGPM, 1983 (*14*)].

The choice of the speed-of-light definition over the use of a particular stabilized laser should ensure that this definition will endure, whereas the krypton definition lasted only 23 years. In practice, a number of ''recommended radiations,'' that is, frequencies of particular stabilized lasers, are published accompanying the definition. This means that to realize the meter to a given uncertainty, it is not necessary to remeasure the frequency of the stabilized laser used.

The differences between the definition of a unit and its realization and practical implementation need to be made clear. To realize the meter, there is no need to measure the distance that light travels in 1/299,792,458 of a second by literally timing a light beam. One can, for example, continue to use a laser interferometer and measure the frequency of the laser used, or use a recommended stabilized laser and then use the relationship $c = f\lambda$ (as well as corrections for refractive index, if the measurement is not done in a vacuum). The realization is a method that implements the definition by using the known laws of physics; it allows the experimental production of a known quantity of the same kind as the one defined, but the method used may be dissimilar to the one in the definition.

Development of Electrical Units

The ampere. The earliest standard to be maintained by means of a fundamental constant was the 1906 international ampere. This standard defined the ampere as current required to deposit 1.18800 mg s^{-1} of silver by electrolysis and hence based the ampere on the Faraday constant, which is given by

$$F = \frac{ItM}{zm} \tag{2}$$

where m is the mass of a substance of molar mass M and valence z deposited by electrolysis using a current I for time t. This ampere definition endured until 1948, the decision to make a change having been postponed by World War II. It was recognized in the 1930s that there were difficulties with the practical implementation; in practice, the ampere was maintained by using standard cells and resistors to maintain the volt and ohm. This situation was resolved by a change in the definition to the modern one: ''The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length'' [9th CGPM, 1948 (*15*)].

This definition is based on the laws of electromagnetism and combined with the expression for the force F per length l on two parallel infinite wires, each carrying a current I, a distance d apart in a vacuum, namely

$$\frac{F}{l} = \frac{\mu_0 I^2}{2\pi d} \tag{3}$$

has the effect of assigning to the value of the magnetic constant μ_0 (also known as the permeability of free space), an exact fixed value of $4\pi \times 10^{-7}$ H m^{-1}. The fixed value of the speed of light means that this definition is equivalent to fixing the value of the electric constant ε_0 (also known as the permittivity of free space) and of the characteristic impedance of free space, through the relationships

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \text{ and } Z_0 = \sqrt{\mu_0/\varepsilon_0} \tag{4}$$

Electrical standards are in practice necessarily derived from mechanical ones. Realization of the ampere by measuring the forces on current-carrying wires is now no longer undertaken, as these experiments were limited at about one part in 10^6 by difficulties in establishing the geometry of the current flow in relation to the wire. Today, the ampere is realized through a combination of realizations of the farad, volt, and watt discussed below.

Absolute realizations of the electrical units. Modern realizations of the electrical units are based on a number of experimental systems that link mechanical and electrical measurement. The ohm can be determined using the calculable capacitor. This is a capacitor whose geometry is such that its capacitance can be accurately calculated using a theorem published in 1956 by Thompson and Lampard (*16*). They found that the cross capacitance of four infinitely long right cylindrical conductors is given by

$$\frac{dC}{dz} = \frac{\varepsilon_0}{\pi}\ln(2)$$

$$= 1.953\ldots \text{pF m}^{-1} \tag{5}$$

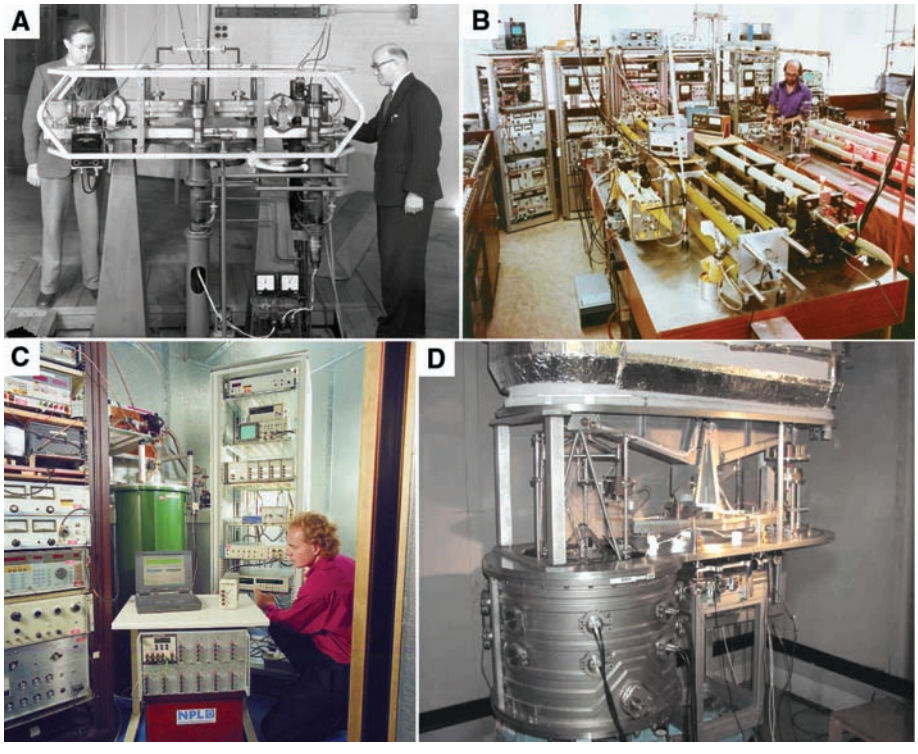


Fig. 1. Some milestones in the development of quantum metrology. (**A**) The cesium beam clock of Essen (1956). (**B**) The speed of light measurement at the National Physical Laboratory (NPL) in the United Kingdom (1979). This experiment and a similar one at the National Bureau of Standards (now NIST) in the United States established the value used in the speed-of-light definition of the meter. (**C**) A modern Josephson volt apparatus. (**D**) The NPL watt balance apparatus.

The only measurement required is the length of the rods making up the capacitor.

The challenge, as with many experiments of this type, is the care needed in construction, measurement, and alignment of a system to the highest possible accuracy. A number of standards laboratories around the world have done experiments of this type and have achieved uncertainties of a few parts in 10^8. Some such measurements are still in progress.

An absolute volt realization was made by Clothier (*17*) at the National Measurement Laboratory, Australia, with an electrometer in which one of the electrodes is a pool of mercury. A vertical electric field, U, is applied, and the change in height of the mercury, d, is measured. The rise of the mercury surface is given by

$$|U| = \sqrt{\frac{2\rho g}{\varepsilon_0 \varepsilon_r}}\, d\, s^{1/2} \qquad (6)$$

where ρ is the density of the mercury, g the acceleration due to gravity, s the interelectrode spacing, and ε_r the relative permittivity of the gas between the electrode and the mercury pool. In practice, the difference between voltages is measured to reduce systematic effects. After considerable effort, this experiment gave a result with an uncertainty of a few parts in 10^7. A capacitance balance measurement at the Physikalisch-Technische Bundesanstalt, Germany (*18*), has achieved a similar accuracy and is in good agreement. It seems unlikely that there will be much progress unless a new method is invented.

Quantum voltage and resistance standards. The start of electrical quantum metrology began with the Josephson effect; indeed, this helped to introduce quantum metrology as a concept. In 1962, Josephson predicted (*19*) that in the presence of an applied microwave field, a direct superconducting tunneling current could pass between superconductors separated by an insulating barrier. This current can only pass when the voltage V across the barrier satisfies the relationship

$$2eV = nh\nu \qquad (7)$$

where e is the electron charge, h the Planck constant, ν the applied frequency, and n an integer. It was recognized that voltage standards could be based on this effect (Fig. 1C). These soon showed that the standard cells used by national laboratories to maintain the volt were drifting at a greater rate than had been believed. A number of experiments found no corrections to expression 7 or dependence on material or experimental conditions at a level of up to parts in 10^{16}. In 1972, a number of countries used the Josephson effect to maintain the volt and agreed on an assigned value for $2e/h$ so that their voltages were in agreement. They are not necessarily the correct SI value; hence, the agreed-upon value is referred to as a "representation" of the volt. Not all countries adopted the same value, but in 1990 international agreement was reached, and the defined value adopted for the frequency-to-voltage conversion $K_{J\text{-}90}$, based on the best knowledge of the volt from absolute volt experiments, was

$$K_{J-90} = 483597.9 \text{ GHz V}^{-1} \qquad (8)$$

This assigned value can be compared with the current best estimate of the value in SI volts; the 2002 Committee on Data for Science and Technology (CODATA) evaluation (*20*) gives

$$K_J = \frac{2e}{h} = K_{J-90}\left[1 - \left(\left(4.3 \pm 8.5\right) \times 10^{-8}\right)\right] \qquad (9)$$

so there is no evidence that the 1990 value is significantly in error.

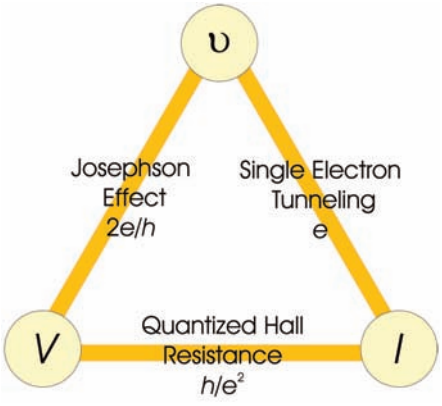


Fig. 2. The metrological triangle. The points of the triangle represent the electrical variables, and the sides represent the experiments embodying quantum effects that link them.

A second electrical quantum effect was demonstrated by von Klitzing in 1980 (*21*), establishing a quantum-based resistance. The quantum Hall effect produces resistance steps in the current-voltage characteristic of a two-dimensional electron gas semiconductor device with values given by

$$R_H = \frac{h}{ie^2} \qquad (10)$$

where h is the Planck constant, e the electron charge, and i an integer index through the steps. Experimental work showed that the resistance of these devices was material and condition-independent to better than one part in 10^9 and, perhaps because of the success of the Josephson effect volt, the quantum Hall resistance standard was quickly adopted, even though the theory behind Eq. 10 is less well established (*22*, *23*). A conventional value of the von Klitzing constant R_{K-90} was adopted in 1990 as a representation of the ohm at the same time as the Josephson constant for the volt (*24*). The conventional value is given by

$$R_{K-90} = 25812.807\ \Omega \qquad (11)$$

Again, this value is based on absolute realizations of the ohm, such as the calculable capacitor, and the CODATA 2002 evaluation gives for the von Klitzing constant

$$R_K = \frac{h}{e^2} = R_{K-90}\left[1 + \left(\left(1.74 \pm 0.33\right) \times 10^{-8}\right)\right] \qquad (12)$$

Fortunately, the small difference between the assigned value and the SI value is inconsequential for most purposes and is well within the uncertainty assigned in 1990.

A further connection enabled by the quantum Hall effect is the link to the fine-structure constant α given by

$$R_K = \frac{h}{e^2} = \frac{\mu_0 c}{2\alpha} \qquad (13)$$

This means that experiments from other parts of physics that determine the fine-structure constant give information on the ohm. The value of the fine-structure constant obtained from the measurement of the anomalous magnetic moment of the electron (*25*) is accurate to 3.8×10^{-9}, whereas the mean of the R_K data gives an uncertainty on the fine-structure constant of 1.8×10^{-8}.

Quantum current standards and the "metrological triangle." The exactness of the expressions $K_J = 2e/h$ and $R_K = h/e^2$ cannot be ensured by theory, so a long-standing goal of electrical metrology is to test the consistency of those relationships using the "metrological triangle" (Fig. 2). The experiment is to compare the voltage produced by a quantized current (controlled passage of counted electrons) passing through a quantum Hall effect resistor with a Josephson voltage. To do this at a meaningful level will require a quantum standard of current accurate at about one part in 10^8, a goal that is actively being pursued in a number of laboratories (*26*). An experiment that demonstrated the principle of electron counting was the charging of a capacitor with electrons counted with a single-electron transistor (*27*), although the rates necessary for closure of the metrological triangle have not been achieved. The present state of the art is about an order of magnitude short of the parts in 10^8 accuracy needed. Electron counting gives a current I comprising electrons in step with a driving frequency ν

$$I = \nu e \qquad (14)$$

This expression combined with expressions 7 and 10 gives a redundancy that enables the exactness of the three relationships to be tested. Although there could still be an offset in all three expressions coming from some as-yet-unknown physics, it would be very surprising if independent systematic errors from three experiments could cancel each other out.

The Kilogram: The Remaining Artifact Base Unit

The kilogram is the remaining artifact standard of the SI (Fig. 3), as is clear from its definition: ''The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram'' [3rd CGPM, 1901 (*28*)].

By definition, the kilogram standard cannot change mass; it is always one kilogram in SI units. However, by considering the physics of the situation, we expect that there is a sense in which its mass could be drifting. Perhaps the intermittent cleaning process to which it is subjected removes some material from the kilogram artifact, or perhaps it absorbs some impurities from the atmosphere. We would consider these mass changes, but this can only be tested experimentally by having some unit of mass that we consider as more stable, for example, a mass based on atomic masses or on the mass equivalent of energy.

Fig. 3. The prototype kilogram. The only object that has a mass of exactly one kilogram. The prototype kilogram, the last artifact unit of the SI, is kept in a vault at the Bureau International des Poids et Mesures in Sèvres near Paris.

Link with electrical unit: the watt balance. Early realizations of the ampere were based on balancing the forces between accurately made coils. In the 1960s, nuclear magnetic resonance (NMR) methods of flux density measurement were introduced, and in 1969 a recommended value of the gyromagnetic ratio of the proton was agreed upon to allow the ampere to be maintained in terms of the flux density produced by a coil. This method was used in standards laboratories to monitor standard cells, which were the practical electrical unit. Until 1990, the unit of flux density (which is volt second per meter squared) disseminated directly by NMR was in closer agreement with the SI than the volt itself could be disseminated via standard cells.

A new method of realizing the watt was proposed by Kibble in 1975 (*29*, *30*), building on the techniques developed for the earlier current measurements with NMR. The method, known as the watt balance, in effect links the mechanical and electrical units of power. Following the discovery of the quantum Hall effect, this can be reinterpreted as a linking of the kilogram to the Planck constant h. In this experiment, a coil is suspended from the arm of a balance in a magnetic field of flux density B, and the current I required to produce a force to balance a mass m is measured. Subsequently, the coil is moved at a measured velocity through the field, and the voltage E generated is measured. These two operations give, for the weighing mode

$$mg = Bk_cI \qquad (15)$$

where k_c is the coil constant, defining its geometrical properties, and g is the acceleration due to gravity. For the moving mode

$$E = Bk_c\frac{dz}{dt} \qquad (16)$$

The elegance of this method is that the geometry factor and the magnetic field can be canceled out, removing the requirement to measure the coil constant that limited ampere balances. Combining Eqs. 15 and 16 gives

$$EI = mg\frac{dz}{dt} \qquad (17)$$

Measuring the mechanical quantities m, g, and dz/dt thus gives an absolute measurement of electrical power EI.

If the voltage is measured in terms of the Josephson volt and the current in terms of the Josephson volt and quantum Hall resistance, the result determines $K_J^2R_K$, which by substitution of the definitions above gives $4/h$. Thus, the left-hand side of Eq. 17 can be considered fixed to the Planck constant, and measuring g and dz/dt gives an electrical method of measuring the kilogram.

Again there are formidable practical difficulties, especially in ensuring that the product Bk_c is the same in both cases. First-generation measurements have been made by the National Physical Laboratory, UK (Fig. 1D), and the National Institute of Standards and Technology (NIST), USA, giving a combined uncertainty of 8.7×10^{-8}. A number of laboratories are now undertaking measurements of this type; these have been reviewed by Eichenberger (*31*). Achieving the target accuracy of one part in 10^8 for the next generation of measurements would allow monitoring of the stability of the prototype kilogram at a significant level compared with the uncertainties associated with its maintenance.

Counting atoms: the Avogadro number method. An alternative to electrical methods for producing a nonartifact mass standard is provided by methods based on counting atomic masses. The most developed of these is the crystallographic method, which is being pursued by an international collaboration (*32*). This method relies on measurements of the lattice spacing of an ultrapure silicon crystal sphere of remarkable roundness. The radius r of the sphere is measured by optical interferometry, and the lattice spacing a determined by x-ray interferometry; hence, the mass m of the crystal will be known in terms of the molar mass of silicon $M(\mathrm{Si})$ and the Avogadro constant N_A, that is

$$m = \frac{nM(\mathrm{Si})\frac{4}{3}\pi r^3}{N_Aa^3} \qquad (18)$$

where n is the number of atoms per unit cell.

The lattice spacing is not a fundamental constant, as it requires standard conditions of pressure and temperature. Considerable effort is needed to determine the measured components in Eq. 18 to the required level of accuracy. Current effort is focused on repeating the experiment, but with a piece of monoisotopic ^{28}Si, because the determination of the isotope ratio is currently one of the principal limiting factors. The silicon sphere will be artifact-like in that it will be difficult to ensure that it remains unchanged and to compare it with secondary standards. Although it could be reproduced if damaged, this would require major and time-consuming effort.

Surprisingly, perhaps, the link between the Avogadro and Planck constants is provided by the Rydberg constant, a constant of atomic spectroscopy given by

$$R_\infty = \frac{m_ec\alpha^2}{2h} \qquad (19)$$

This expression shows the link between the Planck constant and the electron mass m_e. Because atomic mass ratios and the electron-to-proton mass ratios are well known, the electron mass provides a link to the Avogadro. So the Planck constant may be written

$$h = \frac{cA_r(\mathrm{e})M_u\alpha^2}{2R_\infty N_A} \qquad (20)$$

where $A_r(\mathrm{e})$ is the relative atomic mass of the electron and M_u is the molar mass constant, 10^{-3} kg mol^{-1} exactly.

Another less-developed method of this type is also under investigation as a replacement mass standard, an ion accumulation method (*33*). In this method, the current in

an ion beam is accurately measured and the accumulated material is then weighed. An element with only one natural isotope can be used to prevent the isotope ratio problems that have been encountered in the Avogadro method. However, this method is presently limited by the small mass collected and by the difficulty of ensuring that all the ions in the measured current are collected.

A redefinition of the kilogram could be considered based on a fixed value of either the Planck constant or the Avogadro number. In practice, it is not necessary to use the defined method for realization. If, for example, the kilogram were defined by fixing the mass of the carbon-12 atom, which would provide a nice intuitive definition, then in practice the watt balance could still be used to realize the kilogram. Relationships between the fundamental constants, such as Eq. 20, then allow conversions between realization and definition, as well as consistency checks with experiments based on different physical effects.

A speculative pressure method. Future quantum standards based on new or revisited quantum effects will likely take us by surprise, as did some of the effects presently used. One highly speculative possibility that has been considered is an absolute pressure standard based on superfluid helium. Given a measurable area, a pressure standard also provides a force and hence a mass standard.

The Josephson effect in a superconductor that provides the basis of the voltage standards described above has an equivalent in superfluid helium. In this case, the oscillation is in mass flow through a nanoaperture in response to the chemical potential across it. This was proposed as a quantum standard in the 1960s, but the effect could not be observed because of the difficulty of manufacturing apertures small enough to provide the weak link. Recently, the effect has been observed in helium-3 (*34*, *35*) and helium-4 (*36*), with the weak link provided by membranes with arrays of apertures, each on the order of 100 nm in diameter. The Josephson frequency ω_J is given by

$$\omega_J \equiv \frac{\Delta\mu}{\hbar} = \frac{m\Delta P}{\rho\hbar} \tag{21}$$

where ΔP is the pressure difference across the membrane, m the particle mass, and ρ the density of the liquid. In the same form as Eq. 7, this gives

$$\frac{m}{\rho}P = nh\nu \tag{22}$$

To provide a pressure standard, the density of the liquid would have to be measured to high accuracy. The effect is being developed as a candidate rotation sensor, but I am not aware of any standards-focused work that is being undertaken.

Temperature in Terms of Fundamental Constants

Temperature and optical quantities have associated fundamental constants, but the technology that may provide the link allowing the units to be defined in terms of atomic or quantum effects is much further from being established.

Although the kilogram is the last artifact base unit in the SI, the kelvin definition does not meet the criterion of a quantum standard in that it is material dependent and so could be considered "artifact-like." The kelvin is defined as "The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water" [13th CGPM, 1967 (*10*)].

The realization of this definition is limited ultimately in practice by the chemical purity and isotopic content of the water, so that measurable differences exist between triple-point cells from different sources. However, unlike an artifact, a new triple-point cell can be made without reference to another.

To realize the kelvin, one must establish a triple-point equilibrium, but to measure any other temperature it is also necessary to have a primary thermometer to compare the unknown temperature with the temperature of the triple point. A primary thermometer is one whose temperature can be related to physical parameters without unknown constants being introduced; these, therefore, have the potential to link temperature to fundamental constants. Methods of thermodynamic thermometry have been reviewed by Rusby *et al*. (*37*). The most common primary thermometer is the gas thermometer in which the temperature T is related to the pressure p and volume V of a gas of amount of substance n by

$$p = RT\frac{n}{V} \times \left[1 + \frac{n}{V}B(T) + \frac{n^2}{V^2}C(T) + \cdots\right] \tag{23}$$

where R is the gas constant, and $B(T)$, $C(T)$, and so on are virial coefficients that are calculable for a real gas. More precise measurement of the gas constant comes from acoustic thermometry, which measures the speed of sound c and relates this to temperature via the equation of state. In the low pressure limit, this is given by

$$c^2 = \frac{\gamma RT}{M} \tag{24}$$

where γ is the ratio of specific heats and M is the molar mass of the gas but, in practice, similarly to the gas thermometer case, a series of acoustic virial coefficients must be used for a real gas.

A primary thermometer based on a different principle is the total radiation thermometer. Here, the temperature is related to the total radiation of a black body, M, by

$$M = \sigma T^4, \ \sigma = \frac{\pi^2 k^4}{60\hbar^3 c^2} \tag{25}$$

where σ is the Stefan-Boltzmann constant. In this method, the radiation heats a black body absorber, and the heating effect is measured by comparison of the temperature rise it produces to that produced by electrical heating (*38*). This method also provides a link to optical quantities, for if the black body is replaced by a monochromatic light source, then the optical power is measured in terms of electrical power.

Another primary thermometer that has been investigated for metrological use is the noise thermometer, where the mean square electrical noise voltage $\overline{V^2}$ across a resistance R is given by

$$\overline{V^2} = 4kTR\Delta f \tag{26}$$

where Δf is the measurement bandwidth and k the Boltzmann constant. A recent proposition is to use tunnel junction shot noise as a primary thermometer. (*39*).

In all these cases, a fundamental constant relates the measured parameters to temperature; these are conversion factors between thermal and other units. Given the definition of the kelvin as it stands, it is not possible to measure these constants to a greater accuracy than that to which the triple point can be established. If one of these methods were advanced to a stage where it was able to give a resolution better than the stability of the triple-point cell, then it would be appropriate to redefine the kelvin in terms of one of the constants R or k (which are related by the relationship $R = N_A k$). The appropriate primary thermometer would then realize the kelvin and could be used to calibrate fixed points on which a temperature scale would be based. Most likely, a system of fixed points will still be used, as the primary thermometer will not be operable over a wide range of temperatures and is likely to be time consuming and expensive to realize.

It is the unusual properties of the quantity temperature that limit the use of an atomic or quantum definition. The concept of temperature is only applicable to an ensemble. However, under certain conditions, the time series properties of a single state are the same as the properties of an ensemble, and this may pro-

vide a route to a quantized method of temperature measurement.

Conclusions

This short review has summarized thousands of person-years of work linking measurement to atomic and quantum phenomena. As with experiments at the limit of precision, meticulous care is required, and the search for systematic effects seems endless. However, it seems likely that within decades we will have an internationally agreed-upon unit system based entirely on atomic and quantum phenomena, as was envisaged more than a century ago. This will provide a system with a stability and internal consistency based on fundamental constants and, thus, able to provide standards at the low levels of uncertainty required for scientific and technological progress.

Although the route to an SI based on fundamental constants seems nearly complete with the replacement of the artifact kilogram on the horizon, metrology and fundamental constant determinations are far from their ultimate limits.

Both the absolute temperature measurements and the watt balance are unlikely to show the orders-of-magnitude improvement that has been seen with the microwave cesium frequency standard and quantum electrical standards. They are both apparently close to their practical limit already, so any changes to the definitions of the units should be made with caution, as they may not be long-lasting. We await new methods and effects to bring the advances that have been seen in frequency and electrical metrology to other areas of measurement.

References and Notes

1. W. J. Ashworth, *Science* **306**, 1314 (2004).
2. Bureau International des Poids et Mesures, *The International System of Units (SI)*, ed. 7 (1998).
3. Bureau International des Poids et Mesures, *The International System of Units (SI) Supplement 2000: Addenda and Corrigenda to the 7th edition (1998)* (2000).
4. J. C. Maxwell, Address to the Mathematical and Physical Sections of the British Association, Liverpool, September 15, 1870 (from the British Association Report, Vol. XL), reproduced in *The Scientific Papers of James Clerk Maxwell*, W. D. Niven, Ed., Vol. 2 (Cambridge Univ. Press, Cambridge, 1890), p. 225.
5. S. Bize *et al.*, *Phys. Rev. Lett.* **90**, 150802 (2003).
6. M. Fischer *et al.*, *Phys. Rev. Lett.* **92**, 230802 (2004).
7. M. T. Murphy, J. K. Webb, V. V. Flambaum, *Mon. Not. R. Astron. Soc.* **345**, 609 (2003).
8. R. Srianand, H. Chand, P. Petitjean, B. Aracil, *Phys. Rev. Lett.* **92**, 121302 (2004).
9. L. Essen, J. V. L. Parry, *Nature* **176**, 280 (1955).
10. J. Terrien, *Metrologia* **4**, 43 (1968).
11. J. L. Flowers, H. S. Margolis, H. A. Klein, *Contemp. Phys.* **45**, 123 (2004).
12. S. A. Diddams, *Science* **306**, 1318 (2004).
13. B. W. Petley, *Nature* **303**, 373 (1983).
14. P. Giacomo, *Metrologia* **20**, 25 (1984).
15. Procès-Verbaux des Séances du Comité International des Poids et Mesures **20**, 129 (1946).
16. A. M. Thompson, D. G. Lampard, *Nature* **177**, 888 (1956).
17. W. K. Clothier, G. J. Sloggett, H. Bairnsfather, M. F. Curry, D. J. Benjamin, *Metrologia* **26**, 9 (1989).
18. T. Funck, V. Sienknecht, *IEEE Trans. Instrum. Meas.* **40**, 158 (1991).
19. B. D. Josephson, *Phys. Lett.* **1**, 251 (1962).
20. P. J. Mohr, B. N. Taylor, *Rev. Mod. Phys.*, in press; data available at http://physics.nist.gov/cuu/constants (2003).
21. K. von Klitzing, G. Dorda, M. Pepper, *Phys. Rev. Lett.* **45**, 494 (1980).
22. M. M. Nieto, *Phys. Rev. A.* **29**, 3413 (1984).
23. J. E. Avron, D. Osadchy, R. Seiler, *Phys. Today* **56**, 38 (2003).
24. B. N. Taylor, T. J. Witt, *Metrologia* **26**, 47 (1989).
25. R. S. Van Dyck Jr., P. B. Schwinberg, H. G. Dehmel, *Phys. Rev. Lett.* **59**, 26 (1987).
26. F. Piquemal, G. Genevès, *Metrologia* **37**, 207 (2000).
27. M. W. Keller, A. L. Eichenberger, J. M. Martinis, N. M. Zimmerman, *Science* **285**, 1706 (1999).
28. Comptes rendu de la 3e Conférence Général des Poids et Mesures, 70 (1901).
29. B. P. Kibble, in *Atomic Masses and Fundamental Constants 5*, J. H. Sanders, A. H. Wapstra, Eds. (Plenum, London and New York, 1975), pp. 545–551.
30. B. P. Kibble, I. A. Robinson, J. H. Belliss, *Metrologia* **27**, 173 (1990).
31. A. Eichenberger, B. Jeckelmann, P. Richard, *Metrologia* **40**, 356 (2003).
32. P. Becker, *Rep. Prog. Phys.* **64**, 1945 (2001).
33. M. Gläser, *Metrologia* **40**, 376 (2003).
34. S. V. Pereverzev, A. Loshak, S. Backhaus, J. C. Davis, R. E. Packard, *Nature* **388**, 449 (1997).
35. R. W. Simmonds, A. Marchenkov, J. C. Davis, R. E. Packard, *Phys. Rev. Lett.* **87**, 035301 (2001).
36. K. Sukhatme, Y. Mukharsky, T. Chui, D. Pearson, *Nature* **411**, 280 (2001).
37. R. L. Rusby *et al.*, *Metrologia* **33**, 409 (1996).
38. J. E. Martin, N. P. Fox, P. J. Key, *Metrologia* **21**, 147 (1985).
39. L. Spietz, K. W. Lehnert, I. Siddiqi, R. J. Schoelkopf, *Science* **300**, 1929 (2003).
40. Thanks to B. W. Petley, J. C. Gallop, and J.-T. Janssen for useful discussions and suggestions. This work was supported by the UK Department of Trade and Industry Measurement System Policy Unit.

REVIEW

Quantum-Enhanced Measurements: Beating the Standard Quantum Limit

Vittorio Giovannetti,[1] Seth Lloyd,[2]* Lorenzo Maccone[3]

Quantum mechanics, through the Heisenberg uncertainty principle, imposes limits on the precision of measurement. Conventional measurement techniques typically fail to reach these limits. Conventional bounds to the precision of measurements such as the shot noise limit or the standard quantum limit are not as fundamental as the Heisenberg limits and can be beaten using quantum strategies that employ "quantum tricks" such as squeezing and entanglement.

Measurement is a physical process, and the accuracy to which measurements can be performed is governed by the laws of physics. In particular, the behavior of systems at small scales is governed by the laws of quantum mechanics, which place limits on the accuracy to which measurements can be performed. These limits to accuracy take two forms. First, the Heisenberg uncertainty relation (*1*) imposes an intrinsic uncertainty on the values of measurement results of complementary observables such as position and momentum, or the different components of the angular momentum of a rotating object (Fig. 1). Second, every measurement apparatus is itself a quantum system: As a result, the uncertainty relations together with other quantum constraints on the speed of evolution [such as the Margolus-Levitin theorem (*2*)] impose limits on how accurately we can measure quantities, given the amount of physical resources, such as energy, at hand to perform the measurement.

One important consequence of the physical nature of measurement is the so-called quantum back action: The extraction of information from a system can give rise to a feedback effect in which the system configuration after the measurement is determined by the measurement outcome. For example, the most extreme case (the so-called von Neumann or projective measurement) produces a complete determination of the post-measurement state. When performing successive measurements, quantum back action can be detrimental, because earlier measurements can negatively influence successive ones. A common strategy to get around the negative effect of back action

[1]National Enterprise for nanoScience and nanoTechnology–Istituto Nazionale per la Fisica della Materia and Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126, Pisa, Italy. [2]Research Laboratory of Electronics and Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. [3]Quantum Information Theory Group Dipartimento di Fisica "A. Volta," Università di Pavia, via A. Bassi 6 I-27100, Pavia, Italy.

*To whom correspondence should be addressed. E-mail: slloyd@mit.edu

and of Heisenberg uncertainty is to design an experimental apparatus that monitors only one out of a set of incompatible observables: "less is more" (*3*). This strategy, called quantum nondemolition measurement (*3*–*6*), is not as simple as it sounds. One has to account for the system's interaction with the external environment, which tends to extract and disperse information, and for the system dynamics, which can combine the measured observable with incompatible ones. Another strategy to get around the Heisenberg uncertainty is to employ a quantum state in which the uncertainty in the observable to be monitored is very small (at the cost of a very large uncertainty in the complementary observable). The research on quantum-enhanced measurements was spawned by the invention of such techniques (*3*, *7*, *8*) and by the birth of more rigorous treatments of quantum measurements (*9*).

Most standard measurement techniques do not account for these quantum subtleties, so that their precision is limited by otherwise avoidable sources of errors. Typical examples are the environment-induced noise from vacuum fluctuations (the so-called shot noise) that affects the measurement of the electromagnetic field amplitude, and the dynamically induced noise in the position measurement of a free mass [the so-called standard quantum limit (*10*)]. These sources of imprecision are not as fundamental as the unavoidable Heisenberg uncertainty relations, because they originate only from a nonoptimal choice of measurement strategy. However, the shot noise and standard quantum limits set important benchmarks for the quality of a measurement, and they provide an interesting challenge to devise quantum strategies that can defeat them. It is intriguing that almost 30 years after its introduction (*10*), the standard quantum limit has not yet been beaten experimentally in a repeated measurement of a test mass. In the meantime, a paradigm shift has occurred: Quantum mechanics, which used to be just the object of investigation, is now viewed as a tool, a source of exotic and funky effects that can be used to our benefit. In measurement and elsewhere, we are witnessing the birth of quantum technology.

Here we describe some of the techniques that have recently been developed to overcome the limitations of classical measurement strategies. We start with a brief overview of some methods to beat the shot noise limit in interferometry. In the process, we provide a simple example that explains the idea behind many quantum-enhanced measurement strategies. We then give an overview of some of the most promising quantum technology proposals and analyze the standard quantum limit on repeated position measurements. Finally, we show the ultimate resolution achievable in measuring time and space according to the known physical laws. A caveat is in order: This review cannot in any way be viewed as complete, because the improvement of interferometry and measurements through nonclassical light is at the heart of modern quantum optics. Many more ideas and experiments have been devised than can be possibly reported here.

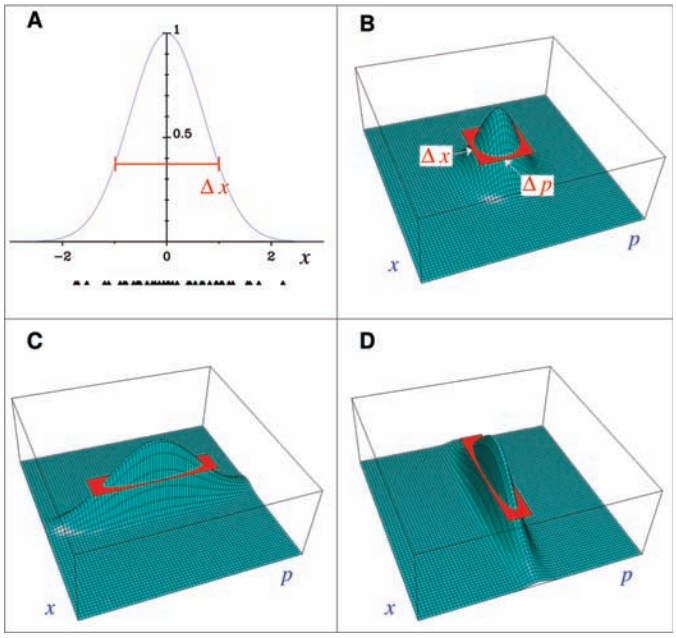


Fig. 1. The Heisenberg uncertainty relation. In quantum mechanics, the outcomes x_1, x_2, etc. of the measurements of a physical quantity x are statistical variables; that is, they are randomly distributed according to a probability determined by the state of the system. A measure of the "sharpness" of a measurement is given by the spread Δx of the outcomes: An example is given in (**A**), where the outcomes (tiny triangles at the bottom of the graph) are distributed according to a Gaussian probability with standard deviation Δx. The Heisenberg uncertainty relation states that when simultaneously measuring incompatible observables such as position x and momentum p, the product of the spreads is lower-bounded: $\Delta x \Delta p \geq \hbar/2$, where $\hbar$ is Planck's constant. The same is true when measuring one of the observables (say x) on a set of particles prepared with a spread Δp on the other observable. [In the general case, when we are measuring two observables A and B the lower bound is given by the expectation value of the commutator between the quantum operators associated to A and B.] In (**B**) we see a coherent state (depicted through its Wigner function). It has the same spreads in position and momentum $\Delta x = \Delta p$. In (**C** and **D**), squeezed states are shown; they have reduced fluctuations in one of the two incompatible observables [x for (C) and p for (D)] at the expense of increased fluctuations in the other. The Heisenberg relation states that the red areas in the plots (given by the product $\Delta x \Delta p$) must have a surface larger than $\hbar/2$. In quantum optics, the observables x and p are replaced by the in-phase and out-of-phase amplitudes of the electromagnetic field; that is, by its "quadratures." Note that the Heisenberg principle is so called only for historical reasons; it is not a principle in modern quantum mechanics, because it is a consequence of the measurement postulate (*1*). Moreover, Heisenberg's formulation of a dynamical disturbance necessarily induced on a system by a measurement was experimentally proven wrong (*81*). It is possible to devise experiments where the disturbance is totally negligible, but where the Heisenberg relations are still valid. They are enforced by the complementarity of quantum mechanics.

Interferometry: Beating the Shot Noise Limit

In this section, we focus on the issues arising in ultraprecise interferometric measurements. A prototypical apparatus is the Mach-Zehnder interferometer (Fig. 2). It acts in the following way. A light beam impinges on a semitransparent mirror (a beam splitter), which divides it into a reflected and a transmitted part. These two components travel along different paths and then are recombined by a second beam splitter. Information on the phase difference φ between the two optical paths of the interferometer can be extracted by monitoring the two output beams, typically by measuring their intensity (the photon number). To see how this works, suppose that a classical coherent beam with N average photons enters the interferometer through the input A. If there is no phase difference φ, all the photons will exit the apparatus at output D. On the other hand, if $\varphi = \pi$ radians, all the photons will exit at output C. In the intermediate situations, a fraction $\cos^2(\varphi/2)$ of the photons will exit at the output D and a fraction $\sin^2(\varphi/2)$ at the output C. By measuring the intensity at the two output ports, one can estimate the value of φ with a statistical error proportional to $1/\sqrt{N}$. This is a consequence of the quantized nature of the electromagnetic field and of the Poissonian statistics of classical light, which in some sense prevents any cooperative behavior among the photons. In fact, the quantity $\cos^2(\varphi/2)$ can be experimentally obtained as the statistical average $\sum_{j=1}^{N} x_j/N$, where x_j takes the value 0 or 1 depending on whether the jth photon in the beam was detected at output C or D, respectively. Because the x_js are independent stochastic variables (photons in the classical beam are uncorrelated), the variance associated with their average is the average of the variances (central limit theorem): The error associated with

the measurement of $\cos^2(\varphi/2)$ is given by $\Delta(\sum_{j=1}^{N} x_j/N) \equiv \sqrt{\sum_{j=1}^{N} \Delta^2 x_j/N} = \Delta x/\sqrt{N}$, where Δx_j is the spread of the jth measurement (the spreads Δx_js are all equal to Δx; they refer to the same experiment). Notice that the same $\sqrt{N}$ dependence can be obtained if, instead of using a classical beam with N average photons, we use N separate single-photon beams. In this case, $\cos^2(\varphi/2)$ is the probability of the photon exiting at output C, and $\sin^2(\varphi/2)$ is the probability of the photon exiting at output D. The $1/\sqrt{N}$ bound on the precision (N being the number of photons used) is referred to as the shot noise limit. It is not fundamental and is only a consequence of the employed classical detection strategy, where neither the state preparation nor the readout takes advantage of quantum correlations.

Carefully designed quantum procedures can beat the $1/\sqrt{N}$ limit. For example, injecting squeezed vacuum in the normally unused port B of the interferometer allows one to achieve a sensitivity of $1/N^{3/4}$ (*7*, *11*). Other strategies can do even better, reaching an $1/N$ sensitivity with a $\sqrt{N}$ improvement over the classical strategies detailed above. The simplest example employs as the input to the interferometer the entangled state (*8*, *12*) $|\Psi\rangle = \frac{1}{\sqrt{2}}(|N_+\rangle_A|N_-\rangle_B + |N_-\rangle_A|N_+\rangle_B)$, where $N_\pm \equiv (N \pm 1)/2$ and where the subscripts A and B label the input ports. This is a highly nonclassical signal, where the correlations between the inputs at A and B cannot be described by a local statistical model. As before, the phase φ can be evaluated by measuring the photon number difference between the two interferometer outputs; that is, by evaluating the expectation value of the operator $M \equiv d^\dagger d - c^\dagger c = (a^\dagger a - b^\dagger b)\cos\varphi + (a^\dagger b + b^\dagger a)\sin\varphi$, where a, b, c, and d are the annihilation operators of the optical modes at the interferometer ports A, B, C, and D respectively (Fig. 2). This scheme allows a sensitivity on the order of $1/N$ for the measurements of small phase differences: $\varphi \cong 0$. In fact, the expectation value of the output photon number difference is equal to $\langle M\rangle = -N_+\sin\varphi$, and its variance is $\Delta^2 M = \cos(2\varphi) + N_+^2 \sin^2\varphi$. The error $\Delta\varphi$ on the estimated phase can be obtained from error propagation, $\Delta\varphi = \Delta M / \left|\frac{\partial\langle M\rangle}{\partial\varphi}\right|$ and for $\varphi \cong 0$, it is easy to see that it scales as $1/N$ (*12*). Even though this procedure achieves good precision only for small values of φ, other schemes exist that show the same high sensitivity for all values of this parameter (*13*). Many quantum procedures that achieve the same $1/N$ sensitivity have been proposed that do not make explicit use of entangled inputs. For example, one can inject squeezed states into both interferometer inputs A and B and then measure the intensity difference at C and D (*14*, *15*), or inject Fock states at A and B and then evaluate the photon-counting probability at the output (*16*), or, finally, measure the de Broglie wavelength of the radiation (*17*). One may wonder whether this $1/N$ precision can be further increased, but in line with the time/energy Heisenberg relation (*18*) and the Margolus-Levitin theorem (*2*), it appears that this is a true quantum limit, and there is no way that it can be beaten (*19*, *20*). It is customarily referred to as the Heisenberg limit to interferometry.

Quantum-Enhanced Parameter Estimation

Some of the above interferometric techniques have also found applications outside the context of optics, such as in spectroscopy (*19*) or in atomic interferometry (*21*). In this section, we point out a general aspect of the quantum estimation theory on which most of the quantum strategies presented in this review are based: the fact that typically a highly correlated input is used and a collective measurement is performed (Fig. 3). A simple example (*22*) may help. Consider a qubit, a two-level quantum system that is described by the two states $|0\rangle$ and $|1\rangle$ and their superpositions. Suppose that the dynamics leaves the state $|0\rangle$ unchanged and adds a phase φ to $|1\rangle$, so that $|1\rangle \rightarrow e^{i\varphi}|1\rangle$. If we want to estimate this phase, we can use a strategy analogous to Ramsey interferometry by preparing the system in the quantum superposition $|\psi_{in}\rangle \equiv (|0\rangle + |1\rangle)/\sqrt{2}$, which is transformed by the system dynamics into $|\psi_{out}\rangle = (|0\rangle + e^{i\varphi}|1\rangle)/\sqrt{2}$. The probability $p(\varphi)$ that the output state $|\psi_{out}\rangle$ is equal to the input $|\psi_{in}\rangle$ allows us to evaluate φ as $p(\varphi) = |\langle\psi_{in}|\psi_{out}\rangle|^2 = \cos^2(\varphi/2)$. This quantity can be estimated with a statistical error $\Delta^2 p(\varphi) = \langle\psi_{out}|(|\psi_{in}\rangle\langle\psi_{in}|)^2|\psi_{out}\rangle - p^2(\varphi) = p(\varphi) - p^2(\varphi)$. If we evaluate a parameter φ from a quantity $p(\varphi)$, error propagation theory tells us that the error associated with the former is given by $\Delta\varphi = \Delta p(\varphi) / \left|\frac{\partial p(\varphi)}{\partial\varphi}\right|$, which in this case gives $\Delta\varphi = 1$. We can improve such an error by repeating the experiment N times. This introduces a factor $1/\sqrt{N}$ in the standard deviation (again as an effect of the central limit theorem), and we find an overall error $\Delta\varphi = 1/\sqrt{N}$. (It is the same sensitivity achieved by the experiment of a single photon in the interferometer described above; these two procedures are essentially equivalent.) As in the case of the interferometer, a more sensitive quantum strategy exists. In fact, instead of using N times the state $|\psi_{in}\rangle$, we can use the following entangled state that still uses N qubits

$$|\phi_{in}\rangle = \frac{1}{\sqrt{2}}(\underbrace{|0\rangle\cdots|0\rangle}_{N \text{ times}} + \underbrace{|1\rangle\cdots|1\rangle}_{N \text{ times}})$$

Now the tensor product structure of quantum mechanics helps us, as the $e^{i\varphi}$ phase factors gained by the $|1\rangle$s combine so that the corresponding output state is

$$|\phi_{out}\rangle = \frac{1}{\sqrt{2}}(|0\rangle\cdots|0\rangle + e^{iN\varphi}|1\rangle\cdots|1\rangle)$$

The probability $q(\varphi)$ that $|\phi_{out}\rangle$ equals $|\phi_{in}\rangle$ is $q(\varphi) = \cos^2(N\varphi/2)$, which, as before, can be estimated with an error $\Delta^2 q(\varphi) = q(\varphi) - q^2(\varphi)$. This means that φ will have an error $\Delta\varphi = \Delta q(\varphi) / \left|\frac{\partial q(\varphi)}{\partial\varphi}\right| = 1/N$. This is a $\sqrt{N}$ enhancement over the precision of N measurements on unentangled qubits, which has been achieved by using an entangled input and performing a collective nonlocal measurement on the output: the measurement of the probability $q(\varphi)$.

A generalization of the parameter estimation presented here is the estimation of the input-output relations of an unknown quantum device. A simple strategy would be to feed the device with a "complete" collection of independent states and measure the resulting outputs. More efficiently, one can use entangled inputs: One-half of the entangled state is fed into the device, and a collective measurement is performed on the other half and on the device's output (*23*, *24*). As in the case discussed above, the quantum correlations between the components of the entangled state increase the precision and hence reduce the number of measurements required. A similar strategy permits us to improve the precision in the estimation of a parameter of an apparatus or to increase the stability of

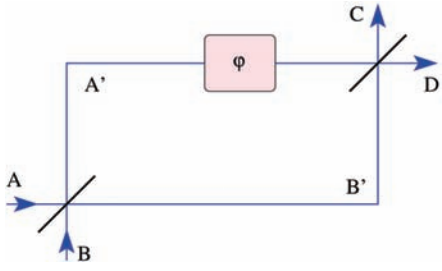


Fig. 2. A Mach-Zehnder interferometer. The light field enters the apparatus through the input ports A and B of the first beam splitter and leaves it through the output ports C and D of the second beam splitter. By measuring the intensities (photon number per second) of the output beams, one can recover the phase difference φ between the two internal optical paths A′ and B′. Formally, the input-output relation of the apparatus is completely characterized by assigning the transformations of the annihilation operators a, b, c, and d associated with the fields at A, B, C, and D, respectively. These are $c \equiv (a' + ie^{i\varphi}|b')/\sqrt{2}$ and $d \equiv (ia' + e^{i\varphi}|b')/\sqrt{2}$, with $a' \equiv (a + ib)/\sqrt{2}$ and $b' \equiv (ia + b)/\sqrt{2}$ the annihilation operators associated with the internal paths A′ and B′, respectively.

measurements (*25*). Part of an entangled state is fed into the apparatus to be probed, and an appropriate collective measurement is performed on the output together with the other part of the entangled state. This permits one, for example, to discriminate among the four Pauli unitary transformations, applying the transformation on only a single qubit probe. It would be impossible without entanglement.

Quantum Technology

The quantum-enhanced parameter estimation presented above has found applications in the most diverse fields. In this section, we give an overview of some of them, leaving aside all the applications that quantum mechanics has found in communication and computation (*26*), which are not directly connected with the subject of this review.

Quantum frequency standards (19, 27). A typical issue in metrology and spectroscopy is to measure time or frequency with very high accuracy. This requires a very precise clock: an oscillator. Atomic transitions are so useful to this aim that the very definition of a second is based on them. To measure time or frequency accurately, we can start with N cold ions in the ground state $|0\rangle$ and apply an electromagnetic pulse that creates independently in each ion an equally weighted superposition $(|0\rangle + |1\rangle)/\sqrt{2}$ of the ground state and of an excited state $|1\rangle$. A subsequent free evolution of the ions for a time t introduces a phase factor between the two states that can be measured at the end of the interval by applying a second, identical electromagnetic pulse and measuring the probability that the final state is $|0\rangle$ (Ramsey interferometry). This procedure is just a physical implementation of the qubit example described above, but here the phase factor is time-dependent and is equal to $\varphi = \omega t$, where ω is the frequency of the transition $|0\rangle \leftrightarrow |1\rangle$. Hence, the same analysis applies: From the N independent ions we can recover the pursued frequency ω (from the phase factor φ) with an error $\Delta\varphi = 1/\sqrt{N}$; that is, $\Delta\omega = 1/(\sqrt{N}t)$.

Instead of acting independently on each ion, one can start from the entangled state $|\phi_{in}\rangle$ introduced above. In this case, the error in the determination of the frequency is $\Delta\omega = 1/(N\ t)$. There is an enhancement of the square root of the number N of entangled ions over the previous strategy.

Quantum lithography and two-photon microscopy (28–32). When we try to resolve objects smaller than the wavelength of the employed light, the wave nature of radiation becomes important, because the light tends to scatter around the object, limiting the achievable resolution. This defines the Rayleigh diffraction bound, which restricts many optical techniques, as it is not always practical to reduce the wavelength. Quantum effects can help by decreasing the wavelength of the light while keeping the wavelength of the radiation field constant. How can this apparently paradoxical effect come about? The basic idea is to use physical devices that are sensitive to the de Broglie wavelength. In quantum mechanics, to every object we can associate wavelength $\lambda = 2\pi\hbar/p$, where p is the object's momentum (for radiation, p is the energy E divided by the speed of light c). Obviously, the wavelength of a single photon $\lambda = 2\pi\hbar c/E = 2\pi c/\omega$ is the wavelength of its radiation field. But what happens if we are able to use a "biphoton" (a single entity constituted by two photons)? In that case, we find that its wavelength is $2\pi\hbar c/(2E) = \lambda/2$: half the wavelength of a single photon or equivalently half the wavelength of its radiation field. Of course, using triphotons, quadriphotons, etc., would result in further decreases of wavelengths. Experimentalists are able to measure the de Broglie wavelengths of biphotons (*17*, *31*, *33*), so that theoreticians have concocted useful ways to employ them. The most important applications are quantum lithography (*28*, *30*), in which smaller wavelengths help to etch smaller integrated circuit elements on a two-photon sensitive substrate; and two-photon microscopy (*32*), in which they produce less damage to the specimens. Also in this context, entanglement is a useful resource because it is instrumental in creating the required biphotons and in enhancing the cross section of two-photon absorption (*29*).

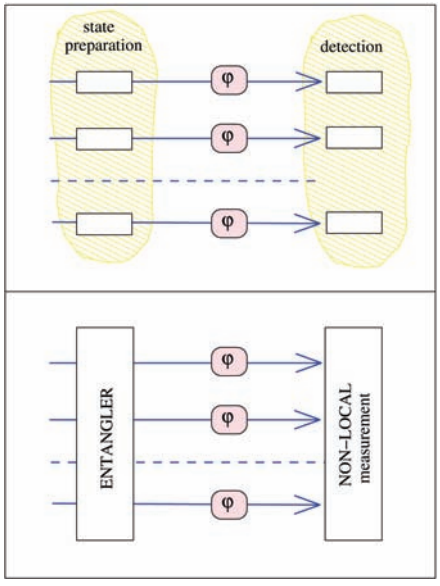


Fig. 3. Comparison between classical and quantum strategies. In conventional measurement schemes (upper panel), N independent physical systems are separately prepared and separately detected. The final result comes from a statistical average of the N outcomes. In quantum-enhanced measurement schemes (lower panel), the N physical systems are typically prepared in a highly correlated configuration (an entangled or a squeezed state) and are measured collectively with a single "nonlocal" measurement that encompasses all the systems.

Quantum positioning and clock synchronization (34–36). To find out the position of an object, one can measure the time it takes for some light signals to travel from that object to some known reference points. The best classical strategy is to measure the travel times of the single photons in the beam and to calculate their average. This allows one to determine the travel time with an error proportional to $1/(\Delta\omega\sqrt{N})$, where $\Delta\omega$ is the signal bandwidth, which induces a minimum time duration of $1/\Delta\omega$ for each photon (that is, the time of arrival of each of the photons will have a spread $1/\Delta\omega$). The accuracy of the travel time measure thus depends on the spectral distribution of the employed signal. The reader will bet that a quantum strategy allows one to do better with the same resources. In fact, by entangling N photons in frequency, we can create a "superphoton" whose bandwidth is still $\Delta\omega$ (it employs the same energetic resources as the N photon signal employed above), but whose mean effective frequency is N times higher, as the entanglement causes the N photons to have the same frequency. This means that the superphoton allows us to achieve N times the accuracy of a single photon with the same bandwidth. To be fair, we need to compare the performance of the superphoton with that of a classical signal of N photons, so that the overall gain of the quantum strategy is $\sqrt{N}$ (*34*).

The problem of localization is intimately connected with the problem of synchronizing distant clocks. In fact, by measuring the time it takes for a signal to travel to known locations, it is possible to synchronize clocks at these locations. This immediately tells us that the above quantum protocol can give a quantum improvement in the precision of distant clocks' synchronization. Moreover, quantum effects can also be useful in avoiding the detrimental effects of dispersion (*37*). The speed of light in dispersive media has a frequency dependence, so that narrow signals (which are constituted by many frequencies) tend to spread out during their travel. This effect ruins the sharp timing signals transmitted. Using the nonlocal correlations of entangled signals, we can engineer frequency-entangled pulses that are not affected by dispersion and that allow clock synchronization (*35*).

Quantum imaging (38, 39). A large number of applications based on the use of quantum effects in spatially multimode light can be grouped under the common label of quantum imaging.

The most famous quantum imaging experiment is the reconstruction of the so-

called ghost images (*40*), where nonlocal correlations between spatially entangled two-photon states are used to create the image of an object without directly looking at it. The basic idea is to illuminate an object with one of the twin photons which is then absorbed by a ''bucket'' detector (it has no spatial resolution, but is able to tell only whether the photon crossed the object or was absorbed by it). The other entangled photon is shone onto an imaging array, and the procedure is repeated many times. Correlating the image on the array with the coincidences between the arrival of one photon at the bucket detector and the other at the imaging array, the shape of the object can be determined. This is equivalent to the following scenario: Use a device that shoots two pebbles in random but exactly opposite directions. When one of the pebbles overshoots an object, it hits a bell. The other pebble instead hits a soft wall, where it remains glued to it. By shooting many pebbles and marking on the wall the pebble's position every time we hear the bell ring, we will project the outline of the object on the wall. The fact that such an intuitive description exists should make us queasy about the real quantum nature of the ghost image experiment, and in fact it was shown that, even if it makes use of highly nonclassical states of light, it is an essentially classical procedure (*41*). The quantum nature of such an experiment lies in the fact that, using the same apparatus, both the near-field and the far-field plane can be perfectly imaged. Classical correlations do not allow this, even though classical thermal light can approximate it (*42*). A related subject is the creation of noiseless images or noiseless image amplification (*38*, *39*), which is the formation of optical images whose amplitude fluctuations are reduced below the shot noise and can be, in principle, suppressed completely.

Many applications require us to measure very accurately the direction in which a focused beam of light is shining. A typical example is atom force microscopy, in which the deflection of a light beam reflected from a cantilever that feels the atomic force can achieve nanometric resolution. Because a light beam is, ultimately, composed of photons, the best way to measure its direction is apparently to shine the beam on an infinitely resolving detector, to measure where each of the photons is inside the beam and to take the average of the positions. This strategy will estimate the position of the beam with an accuracy that scales as $\Delta d/\sqrt{N}$, where Δd is the beam width and N is the number of detected photons. As for the shot noise, this limit derives from the quantized nature of light and from the statistical distribution of the photons inside the beam. As in interferometry, here also quantum effects can boost the sensitivity up to $1/N$ (*11*, *43*, *44*). In fact, consider the following simple example, in which the beam shines along the z direction and is deflected only along the x direction. We can measure such a deflection by shining the beam exactly between two perfectly adjacent detectors and measuring the photon number difference between them. If we expand the spatial modes of the light beam into the sum of an ''even'' mode, which is symmetrical in the x direction, and an ''odd'' mode, which changes sign at $x = 0$ (between the two detectors), we see that the beam is perfectly centered when only the even mode is populated and the odd mode is in the vacuum. Borrowing from sub–shot noise interferometry, we see that we can achieve a $1/N^{3/4}$ sensitivity by populating the odd mode with squeezed vacuum instead. Moreover, we can achieve the Heisenberg limit of $1/N$ by populating both modes with a Fock state $|N/2\rangle$ (*11*).

Any object that creates an image (such as a microscope or a telescope) is necessarily limited by diffraction, because of its finite transverse dimensions. Even though classical ''super-resolution'' techniques are known that can be used to beat the Rayleigh diffraction limit, these are ultimately limited by the quantum fluctuations that introduce undesired quantum noise in the reconstructed image. By illuminating the object with bright multimode squeezed light and by replacing with squeezed vacuum the part that the finite dimensions of the device cuts away, we can increase the resolution of the reconstructed image (*45*), at least in the case of weakly absorbing objects (opaque objects would degrade the squeezed light shining on them).

Coordinate transfer (46–49). A peculiar example of a quantum-enhanced strategy arises in the context of communicating a direction in space (*49*) or a reference frame (*46–48*) (composed by three orthogonal directions). If there is no prior shared reference, it requires some sort of parallel transport such as exchanging gyroscopes (which in quantum mechanical jargon are called spins). Quantum mechanics imposes a bound on the precision with which the axis of a gyroscope can be measured, because the different components of the angular momentum are incompatible observables: Unless one knows the rotation axis a priori, it is impossible to exactly measure the total angular momentum. Gisin and Popescu found the baffling result that sending two gyroscopes pointing in the same direction is less efficient (it allows a less accurate determination of this direction) than sending two gyroscopes pointing in opposite directions (*49*). The reason is that the most efficient measurement for recovering an unknown direction from a couple of spins is an entangled measurement, which has operators with entangled eigenvectors associated to it. Such a detection strategy cannot be separated into different stages, so that it is not possible to rotate the apparatus before the measurement on the second spin, which would imply the equivalence of the two scenarios. The two scenarios could also be shown to be equivalent if it were possible to flip the direction of the second spin without knowing its rotation axis, but this is impossible (it is an anti-unitary transformation, whereas quantum mechanics is notably unitary). Elaborating on this idea, many quantum-enhanced coordinate transfer strategies (*46–48*) have been found.

Repeated Position Measurements: Beating the Standard Quantum Limit

The continuous measure of the position of a free mass is a paradigmatic example of how classical strategies are limited in precision. This experiment is typical of gravitational wave detection, where the position of a test mass must be accurately monitored. The standard quantum limit (*3*, *6*, *10*, *50*) arises in this context by directly applying the Heisenberg relation to two consecutive measurements of the position of the free mass, without taking into account the possibility that the first measurement can be tuned to appropriately change the position configuration of the mass. The original argument was the following: Suppose that we perform the first position measurement at time $t = 0$ with an uncertainty $\Delta x(0)$. This corresponds [via the Heisenberg uncertainty relation (Fig. 1)] to an uncertainty in the initial momentum p at least equal to $\Delta p(0) = \hbar/[2\Delta x(0)]$. The dynamics of an unperturbed the free mass m is governed by the Hamiltonian $H = p^2/2m$, which evolves at time t the position as $x(t) = x(0) + p(0)t/m$. This implies that the uncertainty in the initial momentum $p(0)$ transfers into an uncertainty in the position $x(t)$. The net effect appears to be that a small initial uncertainty $\Delta^2 x(0)$ produces a big final uncertainty $\Delta^2 x(t) \cong \Delta^2 x(0) + \Delta^2 p(0)\, t^2/m^2 \geq 2\Delta x(0)\Delta p(0)\, t/m \geq \hbar t/m$. In this derivation, there is an implicit assumption that the final uncertainty $\Delta x(t)$ cannot be decreased by the correlations between the position and the momentum that build up during the unitary evolution after the first measurement. This is unwarranted: Yuen showed that an exotic detection strategy exists which, after the first measurement, leaves the mass in a ''contractive state'' (*51*); that is, whose position uncertainty decreases for a certain period of time. [The time t for which a mass in such a state has a spread in position $\Delta^2\, x(t)$ below a level $2\delta^2\hbar/m$ satisfies $t \leq 4\delta^2$.] The standard quantum limit is beaten, $\Delta^2 x(t) \leq \hbar/m$, if the second measurement is performed soon enough. The debate then evolved to ascertaining whether two successive measurements at

times 0 and t can be performed, both of which beat the standard quantum limit (*52*). In fact, a simple application of the Heisenberg relation gives $\Delta x(0)\Delta x(t) \geq \frac{1}{2}|\langle[x(0), x(t)]\rangle| = \hbar t/2m$, from which it seems impossible that both measurements have a spread $\leq\sqrt{\hbar t/2m}$. However, $\Delta x(0)$ is the variance of the state immediately after the first measurement, which does not necessarily coincide with the variance of the results of the first measurement. In fact, it is possible (*53*) to measure the position accurately and still leave the mass in a contractive state with initial variance $\Delta x(0) \gg \sqrt{\hbar t/2m}$, so that the standard quantum limit can be beaten repeatedly.

Notice that the back action introduced in the derivation of the standard quantum limit would not occur if one were to measure the momentum instead of the position, because the above Hamiltonian conserves the momentum, $p(t) = p(0)$, which is independent of the position. The momentum measure is an example of a quantum nondemolition detection scheme (*3*–*6*), in which one removes any feedback in the detection by focusing on those observables that are not coupled by the dynamics to their incompatible counterparts.

The standard quantum limit arises also in the context of interferometric measurements of position (*6*, *50*, *54*), where the mass is typically one of the mirrors of the interferometer. The movement of the mirror introduces a phase difference between the arms of the interferometer (Fig. 2). To achieve high measurement precision, one is hence tempted to feed the interferometer electromagnetic signals that possess a well-defined phase. However, the phase and the intensity of the electromagnetic field are in some sense complementary, and a well-defined phase corresponds to a highly undetermined intensity. At first sight this seems without consequences, but any mirror feels a force dependent on the intensity of the light shining on it, through the mechanism of radiation pressure. Hence, the fluctuations in intensity of a signal with a well-defined phase induce a fluctuating random force on the mirror, which ultimately spoils the precision of the measurement setup. Using sufficiently intense coherent light and optimizing the phase and intensity fluctuations, one finds that the attainable precision is again the standard quantum limit (*50*, *54*). Apparently, this derivation of the standard quantum limit is completely independent from the one given above, starting from the Heisenberg relation. However, also here there is an unwarranted assumption: the treatment of phase and intensity fluctuations as independent quantities. Caves showed that by dropping this premise, one can do better (*7*). In fact, a squeezed input signal (Fig. 1) where the amplitude quadrature has less quantum fluctuation than the phase quadrature produces a reduced radiation pressure noise at the expense of an increased photon-counting noise, and vice-versa. This balance allows one to fine-tune the parameters so that the standard quantum limit can be reached with much lower light intensity. Refinements of this technique allows one to beat the standard quantum limit by tailoring appropriate squeezed states (*6*, *14*, *55*–*57*) or by using quantum nondemolition measurements (*3*).

The standard quantum limit is not a fundamental precision threshold. However, at present its conquest is still an open experimental challenge. In fact, on one hand, most of the above theoretical proposals are quite impractical and should be seen only as proofs of principle; and, on the other hand, many competing sources of noise become important when performing very precise measures. The most important is, of course, the thermal fluctuations in the mass to be monitored, but the shot noise at the detection stage or the dissipative part of the mirror response are also big limitations (*3*, *58*–*60*). Various techniques to beat this threshold have been proposed. Among others (by necessity the following list is incomplete), we can cite the use of feedback techniques to enforce a positive back action (*61*–*64*), or the huge number of techniques used to perform quantum nondemolition measurements (*3*, *4*, *65*–*67*), or build contractive states, or build speed meters (*68*). At the present stage, the most promising seem to be the use of nanotechnologies, where tiny mechanical oscillators are coupled to high-sensitivity electronics (*59*, *60*), or the new generations of gravitational wave detectors (*69*).

Quantum Limits to the Measurement of Spacetime Geometry

Quantum effects can be used to increase the accuracy of many different kinds of measurements, but what are the ultimate limits to the resolution that physical laws allow? Attempts to derive quantum limits to the accuracy of measuring the geometry of spacetime date back at least to Wigner (*70*, *71*). As the preceding discussions show, however, care must be taken in applying nonfundamental bounds such as the standard quantum limit. Fortunately, the Margolus-Levitin (*2*) theorem and techniques from the physics of computation (*72*, *73*) can be used to derive limits to the accuracy with which quantum systems can be used to measure spacetime geometry.

The first question is that of minimum distance and time. One can increase the precision of clocks used to measure time by increasing their energy: The Margolus-Levitin theorem implies that the minimum ''tick length'' of a clock with energy E is $\Delta t = \pi\hbar/2E$. Similarly, the wavelength of the particles used to map out space can be decreased by increasing their energy. There appears to be no fundamental physical limit to increasing the energy of the clocks used to measure time and the particles used to measure space, until one reaches the Planck scale, $t_P = \sqrt{\hbar G/c^5} = 5.391\times10^{-44}$s, $l_P = ct_P$. At this scale, the Compton wavelength $2\pi\hbar/mc$ of the clocks and particles is on the same order of magnitude as their Schwarzschild radius $2mG/c^2$, and quantum gravitational effects come into play (*74*). The second question is that of the accuracy to which one can map out the large-scale structure of spacetime. One way to measure the geometry of spacetime is to fill space with a swarm of clocks, all exchanging signals with the other clocks and measuring the signals' times of arrival. In this picture, the clocks could be as large as Global Positioning System satellites or as small as elementary particles.

Let's look at how accurately this swarm of clocks can map out a volume of spacetime with radius R over time T. Every tick of a clock or click of a detector is an elementary event in which a system goes from a state to an orthogonal state. Accordingly, the total number of ticks and clicks that can take place within the volume is a scalar quantity limited by the Margolus-Levitin theorem. It is less than $2ET/\pi\hbar$, where E is the energy of the clocks within the volume.

If we pack the clocks too densely, they will form a black hole and be useless for the measurement of spacetime outside their horizon. To prevent black hole formation, the energy of clocks within a spacelike region of radius R must be less than $Rc^4/2G$. As a result, the total number of elementary events that can occur in the volume of spacetime is no greater than

$$N \equiv \pi^{-1}(T/t_P)(R/l_P) \qquad (1)$$

This quantum geometric limit can also be formulated in a covariant fashion. The maximum number of ticks and clicks in a volume is a scalar quantity proportional to the integral of the trace of the energy-momentum tensor over the four volume; and $2TR$ can be identified with the area of an extremal world sheet contained within the four volume. The quantum geometric limit of Eq. 1 was derived without any recourse to quantum gravity: The Planck scale makes its appearance simply from combining quantum limits to measurement with the requirement that a region not itself be a black hole. (If the region is at or above its critical density, then Eq. 1 still holds if R is the radius of the horizon of the region as measured by an external observer.)

The quantum geometric limit is consistent with and complementary to the Bekenstein bound, the holographic bound, and the covariant entropy bound (*75*–*78*), all of which limit the number of bits that can be contained within a region. [It also confirms

Ng's prediction (*79*) for the scale of spacetime foam.] For example, the argument that leads to Eq. 1 also implies that the maximum number of quanta of wavelength $\lambda \leq 2R$ that can be packed into a volume of radius R without turning that volume into a black hole is bounded by $R^2/\pi l_P^2$ (*80*), in accordance with the Bekenstein bound and holography. Because it bounds the number of elementary events or "ops," rather than the number of bits, the quantum geometric limit of Eq. 1 implies a trade-off between the accuracy with which one can measure time and the accuracy with which one can measure space: The maximum spatial resolution can only be obtained by relaxing the temporal resolution and having each clock tick only once in time T. This lack of temporal resolution is characteristic of systems, such as black holes, that attain the holographic bound (*72*). By contrast, if the events are spread out uniformly in space and time, the number of cells within the spatial volume goes as $(R/l_P)^{3/2}$ (less than the holographic bound), and the number of ticks of each clock over time T goes as $(T/t_P)^{1/2}$. This is the accuracy to which ordinary matter such as radiation and massive particles map out spacetime. Because it is at or close to its critical density, our own universe maps out the geometry of spacetime to an accuracy approaching the absolute limit given by $R^2/\pi l_P^2$: There have been no more than $(T/t_P)^2 \approx 10^{123}$ ticks and clicks since the Big Bang (*73*).

Conclusion

Quantum mechanics governs every aspect of the physical world, including the measuring devices we use to obtain information about that world. Quantum mechanics limits the accuracy of such devices via the Heisenberg uncertainty principle and the Margolus-Levitin theorem, but it also supplies quantum strategies for surpassing semiclassical limits such as the standard quantum limit and the shot noise limit. Starting from strategies to enhance the sensitivity of interferometers and position measurements, scientists and engineers have developed quantum technologies that use effects such as squeezing and entanglement to improve the accuracy of a wide variety of measurements. Some of these quantum techniques are still futuristic; at present, methods for creating and manipulating entangled states are still in their infancy. As we saw, quantum effects usually allow a precision enhancement equal to the square root of the number N of employed particles, but it is usually very complicated to entangle as few as $N = 5$ or 6 particles. In contrast, it is typically rather simple to employ millions of particles to use the classical strategy of plain averaging. As quantum technologies improve, however, the use of entanglement and squeezing to enhance precision measurements is likely to become more widespread. Meanwhile, as the example of quantum limits to measuring spacetime geometry shows, examining the quantum limits to measurement can give insight into the workings of the universe at its most fundamental levels.

References and Notes

1. H. P. Robertson, *Phys. Rev.* **34**, 163 (1929).
2. N. Margolus, L. B. Levitin, *Physica D* **120**, 188 (1998).
3. C. M. Caves, K. S. Thorne, R. W. P. Drever, V. D. Sandberg, M. Zimmermann, *Rev. Mod. Phys.* **52**, 341 (1980).
4. K. Bencheikh, J. A. Levenson, P. Grangier, O. Lopez, *Phys. Rev. Lett.* **75**, 3422 (1995).
5. G. J. Milburn, D. F. Walls, *Phys. Rev. A.* **28**, 2065 (1983).
6. V. B. Braginsky, F. Ya Khalili, *Quantum Measurements* (Cambridge Univ. Press, Cambridge, 1992).
7. C. M. Caves, *Phys. Rev. D* **23**, 1693 (1981).
8. B. Yurke, S. L. McCall, J. R. Klauder, *Phys. Rev. A* **33**, 4033 (1986).
9. C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic Press, New York, 1976).
10. V. B. Braginsky, Y. I. Vorontsov, *Sov. Phys. Usp.* **17**, 644 (1975).
11. S. M. Barnett, C. Fabre, A. Maître, *Eur. Phys. J. D* **22**, 513 (2003).
12. J. P. Dowling, *Phys. Rev. A* **57**, 4736 (1998).
13. B. C. Sanders, G. J. Milburn, *Phys. Rev. Lett.* **75**, 2944 (1995).
14. R. S. Bondurant, J. H. Shapiro, *Phys. Rev. D* **30**, 2548 (1984).
15. M. Xiao, L. Wu, H. J. Kimble Phys, *Rev. Lett.* **59**, 278 (1987).
16. M. J. Holland, K. Burnett, *Phys. Rev. Lett.* **71**, 1355 (1993).
17. J. Jacobson, G. Björk, I. Chuang, Y. Yamamoto, *Phys. Rev. Lett.* **74**, 4835 (1995).
18. L. Mandelstam, I. G. Tamm, *J. Phys. USSR* **9**, 249 (1945).
19. J. J. Bollinger, Wayne M. Itano, D. J. Wineland, D. J. Heinzen, *Phys. Rev. A* **54**, R4649 (1996).
20. Z. Y. Ou, *Phys. Rev. A.* **55**, 2598 (1997).
21. J. Jacobson, G. Bjork, Y. Yamamoto, *Appl. Phys. B* **60**, 187 (1995).
22. P. Kok, S. L. Braunstein, J. P. Dowling, *J. Opt. B Quantum Semiclass. Opt.* **6**, 5811 (2004).
23. G. M. D'Ariano, P. Lo Presti, *Phys. Rev. Lett.* **91**, 047902 (2003).
24. F. De Martini, A. Mazzei, M. Ricci, G. M. D'Ariano, *Phys. Rev. A* **67**, 062307 (2003).
25. G. Mauro D'Ariano, P. Lo Presti, M. G. A. Paris, *Phys. Rev. Lett.* **87**, 270404 (2001).
26. M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
27. S. F. Huelga *et al.*, *Phys. Rev. Lett.* **79**, 3865 (1997).
28. U. W. Rathe, M. O. Scully, *Lett. Math. Phys.* **34**, 297 (1995).
29. H.-B. Fei, B. M. Jost, S. Popescu, B. E. A. Saleh, M. C. Teich, *Phys. Rev. Lett.* **78**, 1679 (1997).
30. A. N. Boto *et al.*, *Phys. Rev. Lett.* **85**, 2733 (2000).
31. M. D'Angelo, M. V. Chekhova, Y. Shih, *Phys. Rev. Lett.* **87**, 013602 (2001).
32. J. N. Gannaway, C. J. R. Sheppard, *Opt. Quantum Electron.* **10**, 435 (1978).
33. E. J. S. Fonseca, C. H. Monken, S. Pádua, *Phys. Rev. Lett.* **82**, 2868 (1999).
34. V. Giovannetti, S. Lloyd, L. Maccone, *Nature* **412**, 417 (2001).
35. V. Giovannetti, S. Lloyd, L. Maccone, F. N. C. Wong, *Phys. Rev. Lett.* **87**, 117902 (2001).
36. R. Jozsa, D. S. Abrams, J. P. Dowling, C. P. Williams, *Phys. Rev. Lett.* **85**, 2010 (2000).
37. J. D. Franson, *Phys. Rev. A* **45**, 3126 (1992).
38. M. I. Kolobov, *Rev. Mod. Phys.* **71**, 1539 (1999).
39. L. A. Lugiato, A. Gatti, E. Brambilla, *J. Opt. B Quantum Semiclass. Opt.* **4**, S176 (2002).
40. T. B. Pittman, Y. H. Shih, D. V. Strekalov, A. V. Sergienko, *Phys. Rev. A* **52**, R3429 (1995).
41. R. S. Bennink, S. J. Bentley, R. W. Boyd, *Phys. Rev. Lett.* **89**, 113601 (2002).
42. A. Gatti, E. Brambilla, M. Bache, and L. A. Lugiato, *Phys. Rev. A* **70**, 013802 (2004).
43. C. Fabre, J. B. Fouet, A. Matre, *Opt. Lett.* **25**, 76 (2000).
44. N. Treps *et al.*, *Phys. Rev. Lett.* **88**, 203601 (2002).
45. M. I. Kolobov, C. Fabre, *Phys. Rev. Lett.* **85**, 3789 (2000).
46. A. Peres, P. F. Scudo, *Phys. Rev. Lett.* **87**, 167901 (2001).
47. E. Bagan, M. Baig, R. Muñoz-Tapia, *Phys. Rev. Lett.* **87**, 257903 (2001).
48. G. Chiribella, G. M. D'Ariano, P. Perinotti, M. F. Sacchi, *Phys. Rev. Lett.* **93**, 180503 (2004).
49. N. Gisin, S. Popescu, *Phys. Rev. Lett.* **83**, 432 (1999).
50. W. A. Edelstein, J. Hough, J. R. Pugh, W. Martin, *J. Phys. E Sci. Instrum.* **11**, 710 (1978).
51. H. P. Yuen, *Phys. Rev. Lett.* **51**, 719 (1983).
52. C. M. Caves, *Phys. Rev. Lett.* **54**, 2465 (1985).
53. M. Ozawa, *Phys. Rev. Lett.* **60**, 385 (1988).
54. C. M. Caves, *Phys. Rev. Lett.* **45**, 75 (1980).
55. W. G. Unruh, in *Quantum Optics, Experimental Gravitation and Measurement Theory*, P. Meystre, M. O. Scully, Eds. (Plenum, New York, 1983), p. 647.
56. M. T. Jaekel, S. Reynaud, *Europhys. Lett.* **13**, 301 (1990).
57. A. F. Pace, M. J. Collett, D. F. Walls, *Phys. Rev. A* **47**, 3173 (1993).
58. M. F. Bocko, R. Onofrio, *Rev. Mod. Phys.* **68**, 755 (1996).
59. R. G. Knobel, A. N. Cleland, *Nature* **424**, 291 (2003).
60. M. D. LaHaye, O. Buu, B. Camarota, K. C. Schwab, *Science* **304**, 74 (2004).
61. H. M. Wiseman, *Phys. Rev. A.* **51**, 2459 (1995).
62. S. Mancini, D. Vitali, P. Tombesi, *Phys. Rev. Lett.* **80**, 688 (1998).
63. P. F. Cohadon, A. Heidmann, M. Pinard, *Phys. Rev. Lett.* **83**, 3174 (1999).
64. G. M. D'Ariano, M. Sacchi, R. Seno, *Nuovo Cimento B* **114**, 775 (1999).
65. N. Imoto, H. A. Haus, Y. Yamamoto, *Phys. Rev. A* **32**, 2287 (1985).
66. G. Nogues *et al.*, *Nature* **400**, 239 (1999).
67. P. Grangier, J. A. Levenson, J.-P. Poizat, *Nature* **396**, 537 (1998).
68. V. B. Braginsky, F. Ya. Khalili, *Phys. Lett. A* **147**, 251 (1990).
69. A. Buonanno, Y. Chen, *Phys. Rev. D* **64**, 042006 (2001).
70. E. P. Wigner, *Rev. Mod. Phys.* **29**, 255 (1957).
71. B. S. De Witt, in *Gravitation: An Introduction to Current Research*, L. Witten, Ed. (Wiley, New York, 1962).
72. S. Lloyd, *Nature* **406**, 1047 (2004).
73. S. Lloyd, *Phys. Rev. Lett.* **88**, 237901 (2002).
74. C. Callender, N. Huggett, Eds., *Physics Meets Philosophy at the Planck Scale: Contemporary Theories in Quantum Gravity* (Cambridge Univ. Press, Cambridge, 2001).
75. J. D. Bekenstein, *Phys. Rev. D* **23**, 287 (1981).
76. G.'t Hooft, in *Basics and Highlights in Fundamental Physics, The Subnuclear Series*, A. Zichichi, Ed. (World Scientific, Erice, Italy, 2001), p. 72 (available at http://lanl.gov/abs/hep-th/?0003004).
77. L. Susskind, *J. Math. Phys.* **36**, 6377 (1995).
78. R. Bousso, *Rev. Mod. Phys.* **74**, 000825 (2002).
79. Y. J. Ng, *Phys. Rev. Lett.* **86**, 2946 (2001) and erratum, Phys. Rev. Lett. **88**, 139902(E) (2002).
80. U. Yurtsever, *Phys. Rev. Lett.* **91**, 41302 (2003).
81. M. O. Scully, B.-G. Englert, H. Walther, *Nature* **351**, 111 (1991).
82. V.G. acknowledges financial support by European Commission (EC) contracts Information Society Technologies–Superconducting Quantum BITS (IST-SQUIBIT), IST-SQUBIT2, and Research Training Networks–Nanoscale Dynamics. S.L. was supported by the Defense Advanced Research Projects Agency, Advanced Research and Development Activity, and Army Research Office via a Multidisciplinary Research Initiative program. L.M. acknowledges financial support by EC Active Teleportation and Entangled State Information Technology project IST-2000-29681 and Ministero dell'Istruzione, dell'Universitá e della Ricerca Cofinanziamento 2003.

The Ocean Takes a Deep Breath

Arne Körtzinger,* Jens Schimanski, Uwe Send, Douglas Wallace

The temperature (T) and salinity (S) of the deep ocean are established at mid- and high latitudes where intermediate and deep waters are formed through subduction and deep convection. These processes also ventilate the deep ocean for atmospheric gases such as oxygen (O_2). Although O_2 is consumed within the ocean by heterotrophic processes, the ocean has no internal oxygen sources. The O_2 concentration in the ocean's interior therefore reflects a balance between supply through circulation and loss through respiration.

We therefore conducted a test deployment of an oxygen sensor mounted on an autonomous profiling float (*5*). Such floats, which report their data by satellite, are being deployed in large numbers (there are presently nearly 1500) to monitor the T and S structure of the oceans. O_2 measurements from such floats could provide tens of thousands of profiles in a single year, a multiple of all data from the unprecedented, 10-year, ship-based World Ocean Circulation Experiment of the 1990s.

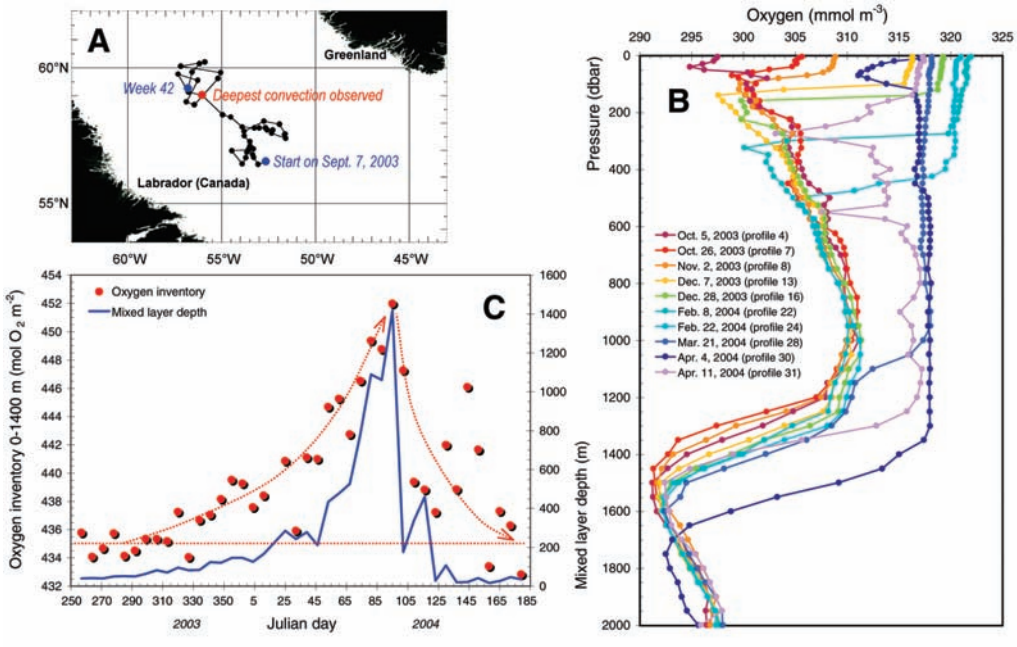


Fig. 1. (**A**) Float track in the central Labrador Sea Gyre, showing positions of weekly surfacing between deployment on 7 September 2003 and profile 42 on 26 June 2004. (**B**) Selected vertical oxygen profiles. (**C**) Temporal development of the oxygen inventory (in the upper 1400 m) and mixed-layer depth, based on 42 weeks of measurements. Dotted red lines represent the build-up and decay of the oxygen inventory in the convection region relative to an assumed background concentration of the surrounding waters.

Long-term trends and variability in atmosphere-ocean O_2 flux complicate the use of atmospheric oxygen time series for global carbon budgeting (*1*). Modeling studies (*2*) have linked flux variations to variability in winter convection. Observations (*3*) of temporal trends in intermediate water O_2 concentrations have been explained by changing ocean ventilation. Models (*4*) also suggest that the air-to-sea O_2 flux will decrease in response to changes in ocean circulation. Monitoring of oceanic oxygen could therefore provide important constraints on the global carbon cycle, as well as a means to monitor changes in ocean circulation.

Our deployment took place in the central Labrador Sea, a region of deep convection subject to strong interannual and interdecadal variability (*6*). The float measured weekly vertical profiles of T, S, and O_2 in the upper 2000 m and remained in the gyre during the fall, winter, and spring (Fig. 1A), i.e., over a full deep convection event. O_2 profiles (Fig. 1B) as well as T and S data (not shown) revealed the transition from a late-summer stratified water column (with a mixed-layer depth of <50 m) into a convectively overturning, deeply mixed late-winter situation (mixed-layer depth of ~1400 m). This change was associated with an inventory increase in the upper 1400 m of 17 mol of O_2 m^{-2} (Fig. 1C), which corresponds to monthly air-to-sea fluxes of up to 7 mol of O_2 m^{-2}. Such fluxes strongly exceed those previously reported for the subpolar North Atlantic (*7*). The oxygen intake was mostly driven by the progressive, almost exponential, deepening of the mixed layer (Fig. 1C), which progressively exposed large volumes of undersaturated water to the atmosphere.

This "deep breath" of a high-latitude deep convection region was observed "live via satellite." The inhalation of oxygen stopped in early April (profile 30), when deep convection ceased. Afterward, the homogeneously mixed volume was rapidly capped by a shallow low-salinity surface layer and sealed from the atmosphere. Subsequently, lateral intrusions into the newly formed water introduced spike-like signatures that likely represent characteristics of water from outside the convection region. The rapid decrease of the O_2 inventory after convection is striking. Outgassing can be ruled out as a major cause. Rather, it appears that the newly added oxygen is rapidly injected into the ocean interior through lateral export of the convectively mixed water and replacement with less oxygenated surrounding waters.

Our results suggest that new observational platforms and sensors could make oxygen a key parameter for addressing major issues of global change research during the 21st century.

References and Notes

1. R. F. Keeling, S. C. Piper, M. Heimann, *Nature* **381**, 218 (1996).
2. G. A. McKinley, M. J. Follows, J. Marshall, *Geophys. Res. Lett.* **27**, 2933 (2000).
3. R. F. Keeling, S. R. Shertz, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7848 (2002).
4. L. Bopp, C. LeQuéré, M. Heimann, A. C. Manning, P. Monfray, *Global Biogeochem. Cycles* **16**, doi: 10.10292001GB001445 (2002).
5. A. Körtzinger, J. Schimanski, U. Send, *J. Atmos. Ocean. Technol.*, in press.
6. R. Curry, M. S. McCartney, *J. Phys. Oceanogr.* **31**, 3374 (2001).
7. H. E. Garcia, R. F. Keeling, *J. Geophys. Res.* **106**, C12, 31155 (2001).
8. Supported by the German Research Foundation through the Sondenforschungsbereich 460.

9 July 2004; accepted 9 September 2004

Leibniz-Institut für Meereswissenschaften, 24105 Kiel, Germany.

*To whom correspondence should be addressed. E-mail: akoertzinger@ifm-geomar.de

Pierolapithecus catalaunicus, a New Middle Miocene Great Ape from Spain

Salvador Moyà-Solà,[1]* Meike Köhler,[1] David M. Alba,[1,2] Isaac Casanovas-Vilar,[1] Jordi Galindo[2]

We describe a partial skeleton with facial cranium of *Pierolapithecus catalaunicus* gen. et sp. nov., a new Middle Miocene (12.5 to 13 million years ago) ape from Barranc de Can Vila 1 (Barcelona, Spain). It is the first known individual of this age that combines well-preserved cranial, dental, and postcranial material. The thorax, lumbar region, and wrist provide evidence of modern ape–like orthograde body design, and the facial morphology includes the basic derived great ape features. The new skeleton reveals that early great apes retained primitive monkeylike characters associated with a derived body structure that permits upright postures of the trunk. *Pierolapithecus*, hence, does not fit the theoretical model that predicts that all characters shared by extant great apes were present in their last common ancestor, but instead points to a large amount of homoplasy in ape evolution. The overall pattern suggests that *Pierolapithecus* is probably close to the last common ancestor of great apes and humans.

Anatomical and molecular data indicate that extant great apes (orangutans, bonobos, common chimpanzees, and gorillas) and humans form a monophyletic group of primates sharing a common ancestor (*1*). Present available evidence based on molecular information suggests that the divergence between great apes and gibbons occurred at the end of the Early Miocene or during the Middle Miocene, depending on the calibration point selected (*1*–*3*). In spite of important recent discoveries and interpretations of a number of Early and Middle Miocene taxa such as *Kenyapithecus* (*4*), *Morotopithecus* (*5*), *Equatorius* (*6*), and *Nacholapithecus* (*7*), the nature of the last common ancestor of extant great apes and humans is in doubt, and the relationships between most Miocene hominoids and extant forms are a matter of ongoing debate (*1*). Here we report the discovery of a Middle Miocene [12.5 to 13 million years ago (Ma)] partial hominoid skeleton at a new locality, Barranc de Can Vila 1 (BCV1) (Hostalets de Pierola, Barcelona, Spain). It is the first known hominoid individual of this stratigraphic age associating well-preserved cranial, dental, and postcranial material. It thus promises to contribute substantially to our understanding of the origin of extant great apes and humans.

[1]Institut de Paleontologia M. Crusafont, Escola Industrial 23, Sabadell, Barcelona 08201, Spain. [2]Palaeotheria, S.C.P. Provençals 5, 2°, 1ª, Barcelona 08019, Spain.

*To whom correspondence should be addressed. E-mail: moyass@diba.es

Systematics. Order Primates Linnaeus, 1758. Suborder Anthropoidea Mivart, 1864. Superfamily Hominoidea Gray, 1825. Family Hominidae Gray, 1825. *Pierolapithecus* gen. nov. **Type species**: *P. catalaunicus* gen. et sp. nov. **Etymology**: Refers to the village close to the site, Els Hostalets de Pierola. **Generic diagnosis**: As for the type species. *P. catalaunicus* gen. et sp. nov. **Holotype:** A partial skeleton of a single adult male individual composed of 83 bones or identifiable fragments [specimen number IPS-21350, housed in the Institut de Paleontologia M. Crusafont (Fig. 1)] with splanchnocranium; left and right maxillae, all cheek teeth, maxillae with both canines and right central incisor, nasals, both zygomatics, lacrimals, and a partial frontal bone; carpals, metacarpals, and several manual phalanges of both hands, tarsals, metatarsals and pedal phalanges, the distal epiphysis of the right patella, the left radius, several long bone diaphyses, two pelvic fragments, three vertebrae, two complete ribs, and 12 other large rib fragments (*8*). **Etymology:** *catalaunicus* refers to Catalunya (Catalonia). **Type locality:** BCV1 (Els Hostalets de Pierola, Barcelona, Spain). **Geological age and stratigraphic position**: The lowermost part of the MN 7/8 bioestratigraphic unit (Middle Miocene, between 12.5 and 13 Ma) (*9*).

Species diagnosis. Face with the frontal processes of the maxillae, the nasals, and the orbits on the same plane. Flat nasals, projecting anteriorly beneath the level of the lower orbital rim. Low face with glabella situated posteriorly (*10*). Thin superciliary arches. Rhinion above P4/ (*10*). High zygomatic root situated anteriorly at the level of M1/ (*10*). High nasoalveolar clivus. Posterior border of the incisive foramen located at the level of P3/. Short, wide, and deep palate. Wide nasal aperture, widest close to the base. Wide interorbital distance. Laterally expanded zygomatics. P3/ of similar size to P4/, with reduced cusp heteromorphy. Elongated molars, except for M3/. Upper molars and premolars without cingula. Peripheral position of lingual cusps in upper molars. Large M2/ and small M3/. Large, low-crowned, and compressed upper canine. Low-crowned and procumbent I1/. Strong rib curvature forming a wide and anteroposteriorly compressed thorax. Large and robust clavicle. Mid-lumbar vertebra lacks ventral keel; robust, wide, and short pedicles of the neural arch; spinous processes oriented slightly caudally; transverse processes insert at the junction between pedicle and body. Transverse process of the last lumbar vertebra arises from the pedicle and is oriented dorsally. Short metacarpals and phalanges. Unfused os centrale. Small triquetrum that does not articulate with ulnar styloid, with crevice for meniscus attachment and pisiform facet shifted distally.

Differential diagnosis. *Pierolapithecus* differs from all known Late Miocene Eurasian hominoids by having a less orthognatous face with a low and posteriorly situated glabella, by the position of the zygomatic root over M1/, by the relative proportions of the upper molars, and by the procumbent I1/. *Pierolapithecus* differs from *Griphopithecus* by the complete lack of cingula, by the long and narrow upper molars, and by the marginal position of lingual cusps on molars. *Pierolapithecus* differs from all known Early and Middle Miocene African hominoids by high zygomatic roots, a wide nasal aperture, a deep palate, flat nasals, the relative proportions of the teeth, and the lack of cingula; and from afropithecids by reduced cusp heteromorphy in the upper premolars.

Description and comparisons: The skull. Considering the stratigraphic age and the geographic location of the new finding, it is of great interest to compare *Pierolapithecus* with the genus *Dryopithecus*, present in the latest Middle and in the Late Miocene of Europe (*11*, *12*). *Pierolapithecus* differs from all species currently included in the genus *Dryopithecus* by the relative size of the upper molars (M2/ > M1/ > M; not known for *D. fontani*), in particular the association of a large M2/ with a small M3/, by the larger, low-crowned, and laterally compressed canines, and the larger, low-crowned, spatulate, and more procumbent central incisor (Table 1). The facial anatomy is completely different

from that of *Dryopithecus*. In lateral view, the face of *Pierolapithecus* (Fig. 2) is low dorsoventrally, with the glabella situated posteriorly at the level of M3/, contrasting with the more orthognathous face of *Dryopithecus*, in which the glabella is situated higher and more anteriorly with respect to the level of the teeth rows (*12*, *13*). *Pierolapithecus* has a relatively shorter and wider palate than *Dryopithecus*, and the zygomatic roots are situated more anteriorly (above M1/) than in *Dryopithecus* (above M2/).

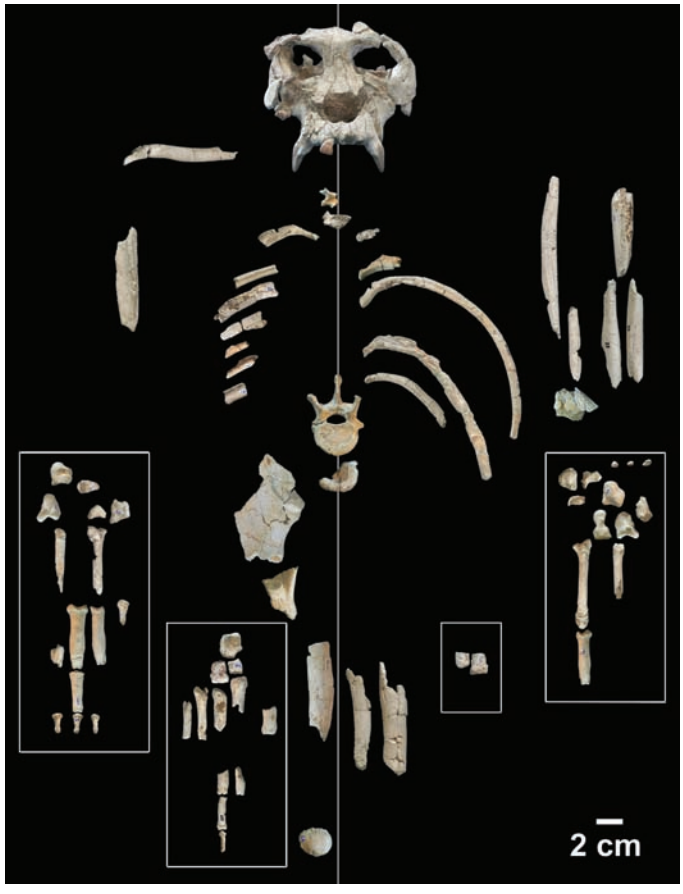


Fig. 1. Complete specimens and large bone fragments of the skeleton of *P. catalaunicus* gen. et sp. nov. (specimen IPS-21350). More fragmentary specimens such as smaller rib fragments, isolated joint facets of vertebrae, and other small specimens are not included.

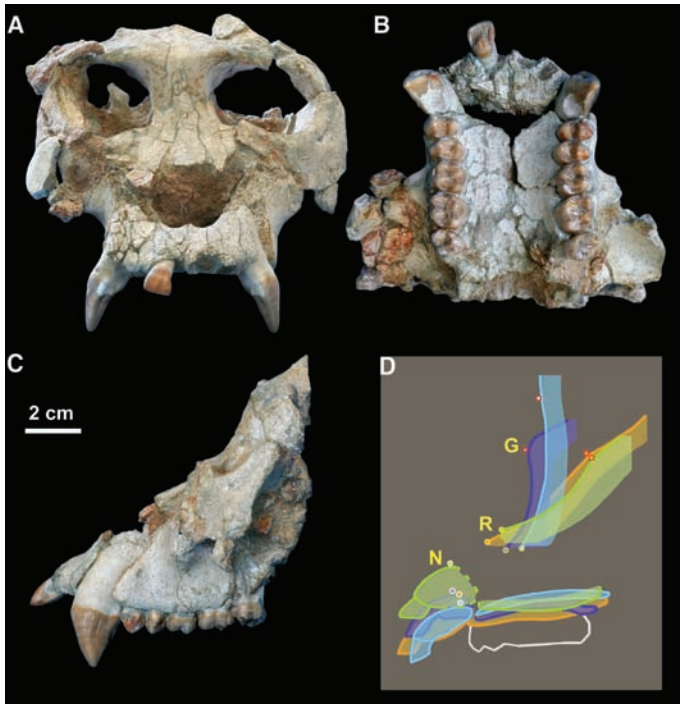


Fig. 2. The face of *Pierolapithecus catalaunicus* (specimen IPS-21350.1). (**A**) Frontal view (the palate is oriented horizontally). (**B**) Palatal view. (**C**) Lateral view. (**D**) Schematic sections of the skulls of *P. catalaunicus* (green), *Afropithecus turkanensis* (orange) from Kalodirr (Kenya) (KNM-WT 16999), *Ouranopithecus macedoniensis* (dark violet) (XIR-1) from Ravin de la Pluie (Greece), and *Sivapithecus indicus* (light blue) (GSP 15000) from the Siwaliks (Pakistan). The sections of the skulls are superimposed and oriented on the base of the cheek teeth drawn to the same length. The primitive profile of *Pierolapithecus* is more similar to *Afropithecus* than to the Eurasian Late Miocene hominoids (here exemplified by *Ouranopithecus* and *Sivapithecus*, two of the best-preserved skulls from the Eurasian Late Miocene), which are characterized by having more orthognathous faces. N, nasospinal; R, rhinion; G, glabella.

In spite of the important differences in facial anatomy, both taxa share a set of derived features such as the frontal processes of the maxillae, the nasals, and the orbits on a same plane; flat nasals that project anteriorly beneath the level of the lower orbital rims, a high zygomatic root, a high nasoalveolar clivus, a deep palate, and a broad nasal aperture widest close to the base. This facial anatomy characterizes extant great apes and must be considered to be shared derived. In the fossil record, these characters are known only in *Pierolapithecus*, *Dryopithecus* (*11*, *12*), *Sivapithecus* (*14*), *Ankarapithecus* (*15*), and *Ouranopithecus* (*16*), but are absent in the known Early and Middle Miocene forms and in gibbons. Thus, *Proconsul* (*17*), *Afropithecus* (*18*), and *Morotopithecus* (*19*) retain a longer muzzle with convex nasals projecting above the lower orbital rim, a primitive catarrhine condition (*20*). Also the known facial specimen of the Middle Miocene hominoid *Nacholapithecus* (*7*, *21*) shows a primitive, *Proconsul*-like, narrow nasal aperture, low zygomatic roots, and a shallow palate.

Although the overall facial structure of *Pierolapithecus* is great ape–like, it still retains a primitive hominoid sagittal profile. The low face with a posteriorly situated glabella and the frontal squama forming an open angle with the orbital plane provide a primitive sagittal profile in which the nasals form an acute angle with the palate (Fig. 2). This contrasts with the more orthognathous profile observed in Late Miocene hominoids and in extant great apes and resembles that of primitive extant hominoids, the hylobatids, or primitive fossil taxa such as *Afropithecus*. This unexpected association of cranial great ape features with a primitive hominoid facial profile is of considerable importance, as it suggests that the facial anatomy of *Pierolapithecus* could make a good prototype for early great apes.

The postcranial skeleton. The partial skeleton of *Pierolapithecus* provides evidence of the morphology of the thorax, the vertebrae, and the carpal-antebrachial complex, key anatomical parts in any diagnostics of the habitual positional and locomotor behaviors that characterize extant hominoids, and thus sources of important phylogenetic information.

Thoracic shape. Differences in thoracic shape between apes and monkeys are significant, with monkeys having a narrow, deep thorax, whereas apes have a broad, shallow thorax, related to an increase in the range of movement of the upper limbs and a shift of the center of gravity in vertical climbing (*22*).

Associated with a broad and shallow thorax in extant hominoids is an increase in rib curvature and angulation. Rib fragments of *Proconsul heseloni* (specimen KNM–RU 2036) (*23*), in particular specimen N°CI, as well as the first rib of *Equatorius africanus* (*6*), suggest a narrow and deep monkeylike thorax for these taxa. The hitherto oldest fossil evidence of apelike thoracic morphology comes from Late Miocene genera [ribs of *Oreopithecus* (*24*, *25*) and thoracic vertebrae of *Dryopithecus* (*26*)]. *Pierolapithecus* now yields the first evidence of a broad and shallow apelike thorax in the Middle Miocene. All complete ribs or rib fragments of *Pierolapithecus* [left rib VIII or IX (specimen IPS 21350-58), proximal right rib III or IV (IPS 21350-66), and proximal left rib XII (IPS 21350-67) (Fig. 3)] show a higher degree of curvature and an increased angulation (Fig. 3) in comparison to the corresponding ribs of monkeys, suggesting a more ventral position of the spinal column in the chest cavity. This constitutes direct evidence that the thorax of *Pierolapithecus* was broader mediolaterally than in monkeys and in known African Miocene hominoids and was close to the pattern of extant apes. This is consistent with a dorsal position of the scapulae on the rib cage inferred from the large size and the chimplike shape of the clavicle (Fig. 1).

The lumbar vertebrae morphology. The morphology of the lumbar vertebrae of extant apes differs considerably from that of monkeys, most noticeably in the shape of the vertebral body, the shape of the pedicle, and in the insertion and orientation of the transverse processes, characters related to a decrease in mobility in the lumbar region of apes (*27*). A nearly complete mid-lumbar vertebra (IPS 21350-64) (Fig. 3) of *Pierolapithecus* resembles the ape pattern in the robustness of the wide and short pedicles, the somewhat caudally oriented spinous process, the reduced wedging, and the lack of the distinct ventral keel and the associated concave shape of the ventrolateral sides found in monkeys and in primitive hominoids such as *Proconsul*. The transverse processes do not arise from the wider part of the vertebral body as in monkeys, *Proconsul* (*27*), and *Nacholapithecus* (*7*), or from the pedicle as in extant great apes, but instead they insert at an intermediate position at the junction between pedicle and body as in extant *Hylobates* and *Ateles*. The position of the transverse processes of the Early Miocene lumbar vertebra of *Morotopithecus* (*27*, *28*) is described by some (*29*) as similar to the pattern of *Pongo*. However, when orienting the specimen with the dorsal surface of the vertebral body (the floor of the neural channel) perpendicular to the lens of the camera, it clearly fits the intermediate pattern of *Hylobates* (*27*) and *Ateles*. In *Pierolapithecus*, the transverse processes of the mid-lumbar vertebrae insert somewhat lower than in *Morotopithecus* (Fig. 4). Among fossil apes, only a mid-lumbar vertebra of *Dryopithecus* unambiguously shows the pattern of *Pongo*, siamangs, and African apes, in which the transverse processes arise directly from the pedicle (Fig. 4). The last lumbar vertebra of *Pierolapithecus* (IPS 21350-65) is most interesting, as the transverse processes are confluent with the pedicle and oriented upwardly, approaching the modern pattern of extant apes (Fig. 3).

The wrist-antebrachial joint. The wrist-antebrachial character complex of extant hominoids is unique among primates and is characterized by a nonarticular ulnar styloid process associated with a semilunar meniscus (*30*, *31*). Functionally, this increases the capacity of adduction (ulnar deviation) and supination at the wrist during climbing and suspension (*30*, *31*). Early Miocene *Proconsul* (*32*) and Middle Miocene *Equatorius* (*6*) are described as having a long styloid process of the ulna in contact with the proximal carpal row as in monkeys, permitting weight transfer in quadrupedal locomotion through both the radius and the ulna. The 11 carpal bones found in Can Vila com-

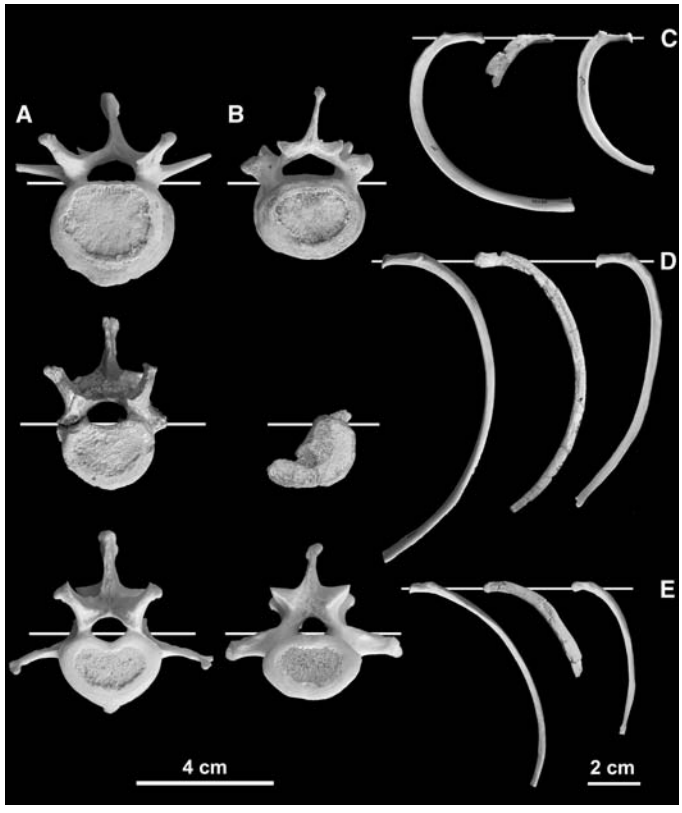


Fig. 3. Ribs and vertebrae of *P. catalaunicus* gen. et sp. nov. (**A**) Mid-lumbar vertebra of *Pan* (top), *P. catalaunicus* (IPS-21350.64) (middle), and *Papio* (bottom). (**B**) Last lumbar vertebra of *Pan* (top), *P. catalaunicus* (IPS-21350.65) (middle), and *Papio* (bottom); scale bar, 4 cm. (**C**) III-IV rib of *Pan* (left), *P. catalaunicus* (IPS-21350.59) (middle), and *Papio* (right). (**D**) VIII-IX rib of *Pan* (left), *P. catalaunicus* (IPS-21350.58) (middle), and *Papio* (right). (**E**) VIII-IX rib of *Pan* (left), *P. catalaunicus* (IPS-21350.67) (middle), and *Papio* (right). The ribs are oriented with the axis of the costal neck parallel to the horizontal white lines.

Table 1. Dental and selected cranial measurements of the face of *P. catalaunicus* (IPS-21350.1).

	Right maxilla		Left maxilla	
	Length (mm)	Breadth (mm)	Length (mm)	Breadth (mm)
M3/	7.65	10.8	8.4	11.7
M2/	11.25	11.9	10.8	12.2
M1/	10.23	11	10	11.3
P4/	7.10	11.8	7.3	11.7
P3/	7.45	11.7	7.64	11.8
C1/	15.6	11.4	16.2	10.9
I1/	7.6	9		
Length from P3/ to M3/		43.8		44
	Cranial measurements (*mm*)			
Maximal interorbital distance		20		
External biorbital breadth (estimated)		110		
Height of glabela from alveolar margin		65		
Breadth of the palate at the level of P3/		29.9		
Breadth of the palate at the level of M2/		27.4		
Height of the zygomatic root		19		
Height of the nasoalveolar clivus		13.7		

prise all carpal elements except for the pisiform. Capitate, hamate, lunate, and triquetrum from the left side articulate perfectly (Fig. 5). In comparison with the other carpals, the compact triquetrum is smaller than in monkeys, *Proconsul* (*32*), and *Hylobates*, but slightly larger than in great apes. It is elongated proximodistally and proximally it has a large, convex, nonarticular summit without any trace of a facet for the styloid process of the ulna. On the distolateral surface, the bone shows a depressed area with small pits and porosity for fibrous attachments homologous to the semilunar crevice of the chimpanzee triquetrum for the attachment of the fibrous capsule of the meniscus. The morphology of *Pierolapithecus* differs, however, from that of *Pan* by its wider and more extended crevice (Fig. 5). The facet for the pisiform is large and flat and is shifted distally in comparison with that of monkeys and *Proconsul* (*31*, *32*). The pisiform facet of *Pierolpithecus* makes an angle of 90° with the lateral surface as typically found in apes, contrasting with the obtuse-angled orientation that occurs in monkeys and *Proconsul* (*32*). The loss of ulnocarpal contact that characterizes extant apes is not known from the fossil record. *Pierolapithecus* thus provides the first evidence that the apelike wrist pattern appeared early in the evolution of apes, dating at least as far back as 13 Ma.

Although *Pierolapithecus* shares the derived wrist-antebrachial morphology with extant hominoids, the hand shows a series of primitive hominoid characters. Some features of the proximal phalanges, such as the proximal articular facet tilted proximodorsally, extending slightly onto the dorsal surface of the shaft; the large and widely separated plantar tubercles that enclose a deep central depression; and a wide and flat proximal articular surface, indicate the use of the hand in palmigrady, a posture rarely adopted by extant apes. The *Pierolapithecus* phalanges (proximal and middle) are less curved and shorter than those of extant apes and Late Miocene *Dryopithecus* species [*D. laietanus* (*26*) and *D. brancoi* (*33*)] (Fig. 4). Taking into account that the body mass of the *Pierolapithecus* specimen is comparable with that of the *D. laietanus* CLl-18800 specimen from Can Llobateres (*34*), and that the hand length/body mass ratio of *Dryopithecus* CLl-18800 fits the pattern of long hands that is typical of extant apes adapted for suspensory behavior (*35*), the considerably shorter phalanges of *Pierolapithecus* rather suggest similarities with the short hand pattern of monkeys (Fig. 4).

Discussion and phylogenetic conclusions. The postcranial skeleton of *Pierolapithecus* preserves key morphological regions that provide evidence for an overall mechanically and functionally congruent modern body structure. Thus, the increased capacity of adduction and supination at the wrist, associated with a wide and anteroposteriorly shallow thorax, the shift of the scapulae onto the back (inferred from the long and chimplike clavicle), and the stiff lumbar region suggest an emphasis on orthograde locomotor and positional behavior. This body structure is diagnostic for extant apes and humans, and little fossil evidence has been documented. The only relatively complete material suggesting this pattern belongs to *Oreopithecus* (*24*, *25*) and, though less complete, to *Dryopithecus* (*26*), both of which are Late Miocene genera. Early and Middle Miocene taxa such as *Proconsul*, *Afropithecus*, *Equatorius*, or *Nacholapithecus*, however, still retain the primitive pronograde monkey-like pattern. Hitherto, the only hint of orthograde skeletal structure in the Early/Middle Miocene comes from a few remains of the axial skeleton of Early Miocene *Morotopithecus* (*5*, *36*). However, the facial skeleton of this genus (*19*) exhibits an overall primitive hominoid pattern, suggesting *Morotopithecus* to be a sister taxon of all extant apes (*6*).

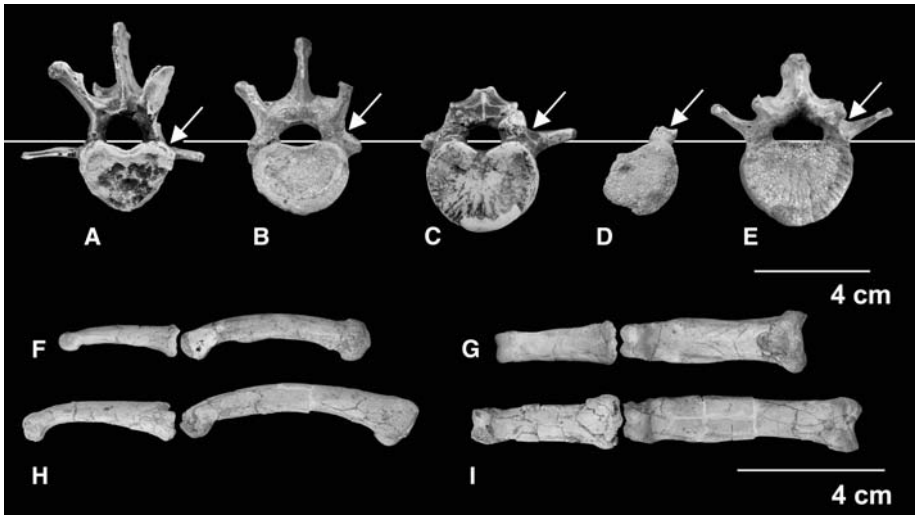


Fig. 4. Comparisons of mid-lumbar vertebrae and medial and proximal phalanges of the third digit. (**A** to **E**) Cranial view of mid-lumbar vertebrae. (A) Cast of *Proconsul nyanzae* (specimen KNM-MW 13142-J) from Mfangano Island (Kenya). (B) *P. catalaunicus* (IPS-21350.59). (C) Cast of *Morotopithecus bishopi* (UPM 67.28) from Moroto (Uganda). (D) *D. laietanus* (IPS-18000) from Can Llobateres (Spain). (E) *Pongo pygmaeus*. Specimens are oriented with the dorsal surface of the vertebral body (the floor of the neural channel) perpendicular to the lens of the camera. The arrows show the origin of the transverse processes. In *Proconsul* (A), the transverse processes arise from the wider part of the vertebral body as in monkeys but do not contact the pedicle. In both *Pierolapithecus* (B) and *Morotopithecus* (C), the transverse processes root partially on the uppermost part of the body and partially on the pedicle. Only in *Dryopithecus* (D) do the transverse processes arise from the pedicle as they do in the extant *Pongo* (E). The white horizontal line represents the dorsal limit of the cranial articular surfaces. (**F** to **I**) Middle and proximal phalanges of the third digit of *Pierolapithecus* compared with those of *D. laietanus* (IPS-18800). (F) and (H) Lateral view. (G) and (I) Palmar view.

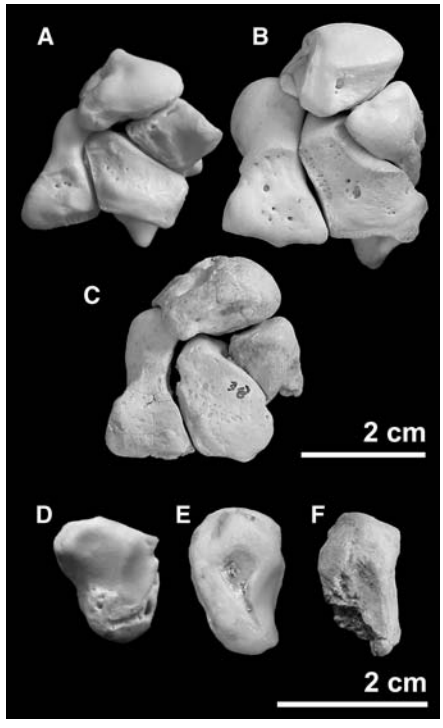


Fig. 5. Comparisons of the carpus of *P. catalaunicus* with that of *Papio* and *Pan*. A dorsal view of the articulated right capitate, hamate, lunate, and triquetrum complex of (**A**) *Papio* sp., (**B**) *Pan troglodytes*, and (**C**) *P. catalaunicus* is shown. Observe the relative large size of the triquetrum and the proximal orientation of the ulnar facet in *Papio*, in comparison with the smaller size and lateral orientation of the triquetrum in *Pan* and *Pierolapithecus*. (**D** to **F**) Lateral view of triquetrum in (D) *Papio* sp., (E) *P. troglodytes*, and (F) *P. catalaunicus*. Note the large articular facet for the ulnar styloid process in *Papio* and its absence in *Pan* and in *Pierolapithecus*, where the homologous surface is nonarticular and shows a crevice for the attachment of a meniscus.

Pierolapithecus, on the contrary, shows the basic derived facial pattern of extant great apes. This facial morphology, combined with the large set of modern ape–like postcranial features, strongly suggests that this taxon is an early member of the great ape and humans clade. The overall pattern suggests that *Pierolapithecus* is close to the last common ancestor of great apes and humans (Fig. 6). Recent molecular results coincide with our interpretation, suggesting that the split between hylobatids and great apes took place about 14.9 ± 2 Ma (*2*) or 14.6 ± 2.6 Ma (*3*).

Nevertheless, the overall postcranial morphology of *Pierolapithecus* is not completely extant ape–like. Although derived features of the chest, lumbar region, and wrist are clearly modern ape–like, *Pierolapithecus* also retains primitive monkeylike skeletal features (such as short phalanges with palmigrade morphological characters) not present in extant apes. This association of primitive and derived features bears important implications for the reconstruction of modern ape evolution. Thus, *Pierolapithecus* provides evidence that the basic orthograde adaptations are not unequivocally and functionally linked to all features commonly shared by extant apes and claimed to be synapomorphies present in the last common ancestor of the group (*1*, *37*). The primitive morphology of the *Pierolapithecus* hand, indicating little (if any) suspensory behavior, strongly suggests that the two basic components of extant ape locomotion—vertical climbing and suspension—appeared independently. Thus, modern ape–like below-branch suspensory locomotion is likely to have been acquired later and independently by the extant members of this clade. Hence, adaptations for below-branch suspensory behaviors might have evolved in parallel and repeatedly, leading to a large amount of homoplasy in ape evolution. This premise suggests that vertical climbing with a basic orthograde body design is the original modern ape adaptation, confirming the hypothesis previously suggested by other authors (*31*, *38*).

The incompleteness of the fossil material of Miocene hominoids, which yields little evidence of the axial skeleton (*1*), and the significant occurrence of homoplasy have combined to obscure the early evolution of great apes, leading to the formulation of different, even contradictory, phylogenetic and taxonomic hypotheses. Our finding provides evidence that the association of a basic orthograde postcranial pattern with a great ape–like facial morphology can be traced back to the Middle Miocene. Under the premise that this association identifies a member of the great ape and humans clade, early great apes are recognizable in the fossil record at least as early as the late Middle Miocene. This new perspective sheds light on two important aspects of hominoid evolution. First, it shows that the known Middle Miocene African taxa are too primitive to be regarded as stem great apes, as has been suggested by some (*39*). Instead, the new information provided by *Pierolapithecus* lends strong support to recent phylogenetic hypotheses (*40*) that exclude the Early/Middle Miocene African taxa with pronograde postcranial pattern and primitive facial anatomy (*Afropithecus*, *Kenyapithecus*, *Equatorius*, and *Nacholapithecus*) from the great ape and humans clade, and classifies them as stem hominoids. Second, our finding provides evidence that early great apes are more primitive than inferred from neontological data (*1*, *37*), because it associates primitive hominoid and derived great ape features. This mosaic of characteristics may explain why most of the known Late Miocene taxa apparently fail to show some of the shared derived features that characterize extant members of the great apes, a perception that has led to ongoing debates (*1*, *12*, *26*, *39–41*) about the high-level phylogenetic relationships between Late Miocene fossil hominoids and extant taxa.

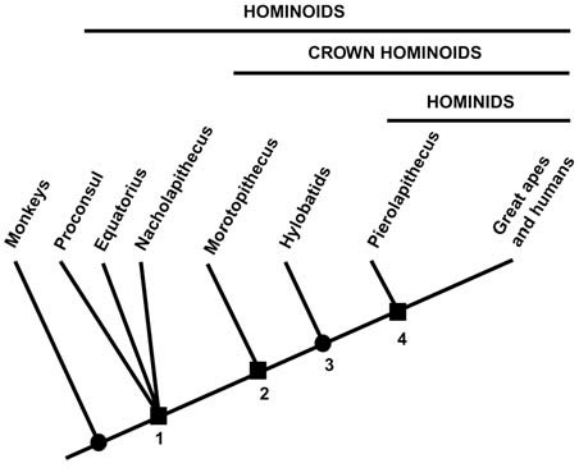


Fig. 6. Cladogram depicting the phylogenetic relationships of Early and Middle Miocene hominoids, with special reference to *P. catalaunicus*. Only those taxa with evidence of the axial skeleton and/or wrist/ulnar joint complex are considered. Therefore, *Afropithecus* does not appear in the cladogram. 1: Absence of tail, medial torsion of the humeral head, low position of the maxillopremaxillary suture/nasal contact, wide anterior palate, and other features discussed in (*39*). 2: The lumbar vertebra of *Morotopithecus* shows clear affinities with extant apes. This is, however, the only evidence that situates this genus within the crown hominoids. 3: Strong rib curvature forming a wide and anteroposteriorly compressed thorax; long and robust clavicle, scapula situated on the back of the thorax; high intermembral index; ulnar shelf of radius excluding ulnar head from joint, mediolaterally broad radiolunate facet; reduced triquetrum with distal migration of pisiform, reduction of the contact between styloid process and triquetrum; lumbar vertebrae reduced in number, lacking ventral keel, with robust, wide, and short pedicles of the neural arch, caudally oriented spinous processes, and transverse processes that do not arise from the broader part of the vertebral body as in monkeys, nor from the pedicle as in extant great apes, but instead insert at an intermediate position at the junction between pedicle and body; wide ilium. 4: The orthograde body plan of *Pierolapithecus* situates this genus within the crown hominoids. The great ape facial anatomy makes it a sister group of great apes and humans. Short face, with the frontal processes of the maxillae, the nasals, and the orbits in the same plane, and flat nasals that project anteriorly beneath the level of the lower orbital rims; high zygomatic roots, a high nasoalveolar clivus, and a broad nasal aperture widest at the base. Triquetrum not articulating with the ulnar styloid. This clade includes all Late Miocene Eurasian hominoids. (Reported characters for this clade are only those that can be observed in *Pierolapithecus* with the available material.)

References and Notes

1. D. Pilbeam, *Mol. Phylogenet. Evol.* **5**, 155 (1996).
2. S. Kumar, B. Hedges, *Nature* **392**, 917 (1998).
3. R. L. Stauffer, A. Walker, O. A. Ryder, M. Lyons-Weiler, S. Hedges, *J. Hered.* **92**, 469 (2001).
4. M. L. McCrossin, B. R. Benefit, in *Function, Phylogeny, and Fossils. Miocene Hominoid Evolution and Adaptations*, D. R. Begun, C. V. Ward, M. D. Rose, Eds. (Plenum, New York, 1997), pp. 225–267.
5. D. L. Gebo *et al.*, *Science* **276**, 401 (1997).
6. S. Ward, B. Brown, A. Hill, J. Kelley, W. Downs, *Science* **285**, 1382 (1999).
7. H. Ishida, Y. Kunimatsu, T. Takano, Y. Nakano, M. Nakatsukasa, *J. Hum. Evol.* **48**, 69 (2003).
8. Two lower molars discovered in 1941 (*42*) in the same area (between Can Vila and Can Mata) were described as a new species of the genus *Sivapithecus*, *S. occidentalis* (*43*). Later this species was synonymized with *Dryopithecus* (*Hispanopithecus*) *laietanus* (*44*), an attribution that has been generally accepted. Considering that the two molars are undistinguishable from those of *D. laietanus* and that neither the site (Can Vila) nor the stratigraphic horizon is identifiable, because the area is larger than 2 km^2 and the stratigraphic sequence is 500 m thick, spanning biozones MN 6 through MN 7/8 (Middle Miocene), the species name *occidentalis* cannot be applied to the new genus and must be considered a nomen dubium (*45*).
9. Fauna and age of BCV1. The fauna recovered in BCV1 comprises 17 mammalian species. The rich rodent fauna is dominated by cricetids (mainly *Eumyarion* and *Democricetodon*) and also includes several glirids, ground and flying squirrels. This faunal association and the relative abundance of its representatives suggests the existence of a wooded and relatively humid biotope, similar to other localities of the Late Aragonian such as Sant Quirze A. On the basis of the presence of *Democricetodon larteti,* this site is placed in MN 7/8. However, *Democricetodon larteti* from BCV1 is very similar in size and morphology to younger populations of species from the Calatayud-Teruel Basin. The age of these "advanced" populations of the Calatayud-Teruel area is close to 12.5 Ma (*46*, *47*). Thus, on the basis of biostratigraphic data, BCV1 is correlated to the lowermost part of the MN 7/8 zone, between 12.5 and 13 Ma (*48*). The faunal list is as follows: *Democricetodon larteti*, *Democricetodon gaillardi*, *Eumyarion* aff. *leemani*, *Megacricetodon* minor *debruijni*,

Cricetodon sp., *Microdyromys complicatus*, *Paraglirulus werenfelsi*, *Glirudinus undosus*, *Muscardinus sansaniensis*, *Bransatoglis* sp., *Spermophilinus bredai*, *Albanensia albanensis*, Soricidae indet., Erinaceidae indet., *Deinotherium giganteum*, *Euprox furcatus*, *Dorcatherium* sp., *Listriodon splendens*, Carnivora indet.
10. These descriptions are made with the tooth row oriented horizontally.
11. D. Begun, *Yrbk. Phys. Anthropol.* **37**, 11 (1994).
12. S. Moyà Solà, M. Köhler, D.M. Alba, in *Hominoid Evolution and Climatic Change in Europe; Vol. 2: Phylogeny of the Neogene Hominoid Primates of Eurasia*, L. de Bonis, G. K. Koufos, P. Andrews, Eds. (Cambridge Univ. Press, Cambridge, 2001), pp. 192–215.
13. L. Kordos, D. R. Begun, *J. Hum. Evol.* **41**, 689 (2002).
14. D. Pilbeam, *Nature* **295**, 232 (1982).
15. B. Alpagut *et al.*, *Nature* **382**, 349 (1996).
16. L. de Bonis, G. Bouvrain, D. Geraads, G. Koufos, *Nature* **345**, 712 (1990).
17. W. E. Le Gros Clark, L. Leakey, *Fossil Mamm. Afr.* **1**, 1 (1951).
18. R. E. Leakey, M. G. Leakey, *Nature* **324**, 143 (1986).
19. M. Pickford, *Hum. Evol.* **17**, 1 (2002).
20. M. G. Leakey, R. E. Leakey, J. T. Richtsmeier, E. L. Simons, A. C. Walker, *Folia Primatol.* **28**, 519 (1991).
21. Y. Kunimatsu *et al.*, *J. Hum. Evol.* **46**, 365 (2004).
22. A. H. Schultz, *Primatology* **4**, 1 (1961).
23. J. Napier, P. R. Davis, *Br. Mus. Nat. Hist. Fossil Mamm. Afr.* **16**, 1 (1959).
24. A. H. Schultz, *Z. Morph. Anthropol.* **50**, 136 (1960).
25. T. Harrison, *J. Hum. Evol.* **15**, 541 (1987).
26. S. Moyà-Solà, M. Köhler, *Nature* **379**, 156 (1996).
27. C. V. Ward, *Am. J. Phys. Anthropol.* **92**, 291 (1993).
28. A. Walker, M. D. Rose, *Nature* **217**, 980 (1968).
29. W. J. Sanders, B. E. Bodenbender, *J. Hum. Evol.* **26**, 203 (1993).
30. O. J. Lewis, in *Primate Locomotion*, F. A. Jenkins, Ed. (Academic Press, New York, 1974), pp. 143–169.
31. E. Sarmiento, *Int. J. Primatol.* **9**, 281 (1988).
32. C. Beard, M. F. Teaford, M. Walker, *Folia Primatol.* **47**, 97 (1986).
33. D. R. Begun, *J. Hum. Evol.* **24**, 737 (1993).
34. Body mass (BM) estimates were based on the regressions (*29*) derived for the following measurements, taken on the mid-lumbar vertebra (level L VI): vertebral body width at the cranial end = 30.0 mm, vertebral body height at the cranial end = 21.7 mm, caudal surface area of the vertebral body = 5.7 cm^2 for nonhuman catarrhines, and vertebral body length at the ventral margin = 22.0 mm for nonhuman hominoids. BM estimates derived from these measurements (23, 32, 24 and 42 kg, respectively) give an average value for the body mass of the *Pierolapithecus* specimen of about 30 kg. Results from dental parameters suggest that the new taxon would be approximately of the same size as IPS18000 *D. laietanus* from Can Llobateres, because both provide the same postcanine tooth row length. BM estimation from M1/ area (averaging left and right measurements) on the basis of a regression equation for males and females separately (*49*, *50*) yields a value of 32 kg for *Pierolapithecus*, which is somewhat larger than the 29 kg obtained for the IPS18800 *Dryopithecus* specimen. 34 kg was obtained for specimen IPS18000 on the basis of femoral head estimators (*26*), a parameter not available for the BCV1 skeleton. For this individual, we thus propose a BM between 30 and 35 kg, which is very similar to that of the *D. laietanus* male skeleton of Can Llobateres (Spain) (*26*).
35. S. Moyà-Solà, M. Köhler, L. Rook, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 313 (1999).
36. N. M. Young, L. MacLatchy, *J. Hum. Evol.* **46**, 163 (2003).
37. T. Harrison, *J. Hum. Evol.* **16**, 41 (1987).
38. J. Fleagle, *Folia Primatol.* **26**, 245 (1976).
39. P. Andrews, *Nature* **360**, 641 (1992).
40. D. R. Begun, C. V. Ward, M. D. Rose, in *Function, Phylogeny, and Fossils. Miocene Hominoid Evolution and Adaptations*, D. R. Begun, C. V. Ward, M. D. Rose, Eds. (Plenum, New York, 1997), pp. 389–415.
41. D. Pilbeam, in *Function, Phylogeny, and Fossils. Miocene Hominoid Evolution and Adaptations*, D. R. Begun, C. V. Ward, M. D. Rose, Eds. (Plenum, New York, 1997), pp. 13–28.
42. J. F. Villalta, M. Crusafont, *Bol. Inst. Geol. Min. Esp.* **55**, 129 (1941).
43. J. F. Villalta, M. Crusafont, *Not. Com. Inst. Geol Min. Esp.* **13**, 1 (1944).
44. M. Crusafont, J. Hürzeler, *C. R. Acad. Sci. Paris* **252**, 562 (1961).
45. A. Mones, *J. Vert. Paleontol.* **9**, 2, 232 (1989).
46. R. Daams, A. J. van der Meulen, M. A. Alvarez-Sierra, P. Peláez-Campomanes, W. Krijgsman, *Earth Planet. Sci. Lett.* **165**, 287 (1999).
47. A. J. van der Meulen, P. Peláez-Campomanes, R. Daams, *Coloquios Paleontol.* **V.E.1**, 385 (2003).
48. J. Agustí *et al.*, *Earth Sci. Rev.* **52**, 247 (2001).
49. P. D. Gingerich, B. H. Smith, K. Rosenberg, *Am. J. Phys. Anthropol.* **58**, 81 (1982).
50. R. J. Smith, W. L. Jungers, *J. Hum. Evol.* **32**, 523 (1997).
51. We thank M. Brunet, F. K. Howell, B. Senut, M. Pickford, D. Pilbeam, L. Rook, T. D. White, and two anonymous referees for comments on the manuscript and for improving style and spelling. We thank P. Andrews for casts of Pasalar specimens. This study has been supported by the Diputació de Barcelona, Departaments d'Universitats Recerca i Societat de l'Informació (grant 2003 FI 00083) and Cultura de la Generalitat de Catalunya, Cespa Gestión de Residuos, Ministerio de Ciencia y Technología (project no. BTE2001-1076), Fundació La Caixa and Fundación Conjunto Paleontológico de Teruel. The support of the Researching Hominid Origins Initiative (RHOI-HOMINID-NSF-BCS-0321893) is gratefully acknowledged. We also acknowledge the collaboration of the Ajuntament dels Hostalets de Pierola. We thank I. Pellejero and S. Val for the excellent restoration of the specimens and À. Blanco, L. Checa, C. Rotgers, and B. Poza for their enthusiasm and collaboration during excavation. We thank W. Kelson for improving the English. The first two authors dedicate this work to the memory of the late J. Pons, an enthusiastic paleontologist and good friend.

22 July 2004; accepted 18 October 2004

The 1.2-Megabase Genome Sequence of Mimivirus

Didier Raoult,[1]* Stéphane Audic,[2] Catherine Robert,[1] Chantal Abergel,[2] Patricia Renesto,[1] Hiroyuki Ogata,[2] Bernard La Scola,[1] Marie Suzan,[1] Jean-Michel Claverie[2]*

We recently reported the discovery and preliminary characterization of Mimivirus, the largest known virus, with a 400-nanometer particle size comparable to mycoplasma. Mimivirus is a double-stranded DNA virus growing in amoebae. We now present its 1,181,404–base pair genome sequence, consisting of 1262 putative open reading frames, 10% of which exhibit a similarity to proteins of known functions. In addition to exceptional genome size, Mimivirus exhibits many features that distinguish it from other nucleocytoplasmic large DNA viruses. The most unexpected is the presence of numerous genes encoding central protein-translation components, including four amino-acyl transfer RNA synthetases, peptide release factor 1, translation elongation factor EF-TU, and translation initiation factor 1. The genome also exhibits six tRNAs. Other notable features include the presence of both type I and type II topoisomerases, components of all DNA repair pathways, many polysaccharide synthesis enzymes, and one intein-containing gene. The size and complexity of the Mimivirus genome challenge the established frontier between viruses and parasitic cellular organisms. This new sequence data might help shed a new light on the origin of DNA viruses and their role in the early evolution of eukaryotes.

Mimivirus, the sole member of the newly proposed *Mimiviridae* family of nucleocytoplasmic large DNA viruses (NCLDVs) was recently isolated from amoebae growing in the water of a cooling tower of a hospital in Bradford, England, in the context of pneumonia outbreak (*1*). The study of Mimivirus grown in *Acanthamoeba polyphaga* revealed a mature particle with the characteristic morphology of an icosahedral capsid with a diameter of at least 400 nm. Such a virion size comparable to that of a mycoplasma cell makes Mimivirus the largest virus identified so far. A phylogenetic study with preliminary sequence data from a handful of conserved viral genes tentatively classified Mimivirus in a new independent branch of NCLDVs (*1*). The sequencing of the genome of Mimivirus was undertaken to determine its complete gene content, to predict some of its physiology, to confirm its phylogenetic position among known viruses, and to gain insight on the origin of NCLDVs.

[1]Unité des Rickettsies, Faculté de Médecine, CNRS UMR6020, Université de la Méditerranée, 13385 Marseille Cedex 05, France. [2]Information Génomique et Structurale (IGS), CNRS UPR2589, Institut de Biologie Structurale et Microbiologie, 13402 Marseille Cedex 20, France.

*To whom correspondence should be addressed. E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr (J.-M.C.); Didier.Raoult@medecine.univ-mrs.fr (D.R.)

Overall Genome Structure

The Mimivirus genome (Fig. 1) was assembled (*2*) into a contiguous linear sequence of 1,181,404 base pairs (bp), significantly larger than our initial conservative estimate of 800 kbp (*1*). The size and linear structure of the genome were confirmed by restriction digests and pulsed-field gel electrophoresis. Two inverted repeats of about 900 nucleotides are

found near both extremities of the assembled sequence, suggesting that the Mimivirus genome might adopt a circular topology as a result of their annealing, as in some other NCLDVs. From transmission electron microscopy pictures, we estimated the volume of the dark central core of the virion (approximated as a sphere) at about 2.6×10^{-21} m^3, which is 3.7 times as large as the core volume of *Paramecium bursaria* chlorella virus (PBCV-1) (*3*). This is quite consistent with the respective genome sizes (1180/331 kb = 3.56) of the two viruses, indicating similar physical constraints for DNA packing (i.e., a core DNA concentration of about 450 mg/ml).

The nucleotide composition was 72.0% A+T. The genome exhibited a significant strand asymmetry. Both the cumulative A+C excess and the cumulative gene excess plots (*2*) (fig. S1) exhibit a slope reversal (around position 400,000, Fig. 1) as found in bacterial genomes and usually associated with the location of the origin of replication. Mimivirus genes are preferentially transcribed away from this putative origin of replication. Despite this local asymmetry, the total numbers of genes transcribed from either strand are similar [450 "R" versus 461 "L" open reading frames (ORFs)]. Repeated sequences represented less than 2.2% of the Mimivirus genome (*2*).

We identified a total of 1262 putative ORFs of length ≥100 amino acid residues, corresponding to a theoretical coding density of 90.5%. Of these ORFs, 911 were predicted to be protein-coding genes, based on their statistical coding propensity and/or their similarity to database sequences. The remaining ORFs have been downgraded to the unidentified reading frame category. We were able to associate 298 ORFs with functional attributes (*2*).

The overall amino acid composition of the predicted Mimivirus proteome exhibits a strong positive bias for residues encoded by codons rich in A+T. For instance, isoleucine (9.87%), asparagine (8.89%), and tyrosine (5.43%) are twice as frequent in Mimivirus than in amoeba or human proteins. Alanine (encoded by A+T-poor codons GCN) is half as frequent (3.06%) as in the other two organisms. Similar variations have been observed in the amino acid compositions of other DNA viruses rich in A+T (*4*). For any given amino acid, the relative usage of synonymous codons is also biased by the A+T-rich genome composition. For instance, ATT is largely dominant for Ile, as is AAT for Asn and TAT for tyrosine. In contrast, GCG is rarely used for Ala, CGG is rarely used for Arg, and GGG and GGC are rarely used for Gly. The codon usage in Mimivirus is almost the exact opposite of the one exhibited by *Acanthamoeba castellanii*: The least frequent codon in the amoeba is systematically the dominant one for Mimivirus. The codon usage in human genes also differs from the one in Mimivirus but to a lesser extent because of the more even vertebrate codon distribution.

NCLDV Core Genes Identified in the Mimivirus Genome

Iyer *et al*. (*5*) identified a set of genes present in all or most members of the four main NCLDV families: *Poxviridae*, *Phycodnaviridae*, *Asfarviridae*, and *Iridoviridae*. These core genes are subdivided into four classes, from the most to least evolutionarily conserved: Class I includes those found in all known NCLDV genome sequences, class II genes are found in all NCLDV clades but are missing in some species; class III genes are identified in three out of the four NCLDV clades; and class IV genes are found in two clades only (*5*). The pattern of presence and absence of Class I, II, and III core genes in Mimivirus is summarized in Table 1. We identified homologs for all (9 out of 9) class I genes, 6 out of 8 class II genes, 11 out of 14 class III genes, and 16 out of 30 class IV genes (*2*) (table S2). Both class II genes that are missing in Mimivirus are relevant to the biosynthesis of 3′-deoxythymidine 5′-triphosphate: thymidylate kinase and 3′-deoxipyridine-5′triphosphate pyrophosphatase (dUTPase), a paradox given its A+T-rich genome. *Ectocarpus silicosus* virus (ESV) also lacks these enzymes. However, Mimivirus exhibits homologs for the class IV core genes thymidylate synthase and thymidine kinase. Additional nucleotide synthesis enzymes include deoxynucleoside kinase (DNK) and cytidine deaminase, as well as the first nucleoside diphosphate kinase (NDK) identified in a double-stranded DNA (dsDNA) virus. Mimivirus also lacks an adenosine 5′-triphosphate (ATP)–dependent DNA ligase (a class III core gene), which was apparently replaced by a nicotinamide adenine dinucleotide (NAD)–dependent ATP ligase (class IV), as found in Iridoviruses (*5*). With the exception of RNA polymerase subunit 10, the Mimivirus genome exhibits the same transcription-related core genes as found in *Poxviridae* and *Asfarviridae*. This suggests that the transcription of at least some Mimivirus genes occurs in the cytoplasm. Overall, the pattern of presence and absence of core genes (class II to IV) in Mimivirus is unlike any of the established patterns. This confirms our initial suggestion (*1*) that Mimivirus constitutes the first representative of a new distinct NCLDV class (the "*Mimiviridae*").

Global Gene Content Statistics

All predicted Mimivirus ORFs were compared with the Clusters of Orthologous Groups (COG) database (*6*) with the Reverse PSI-BLAST

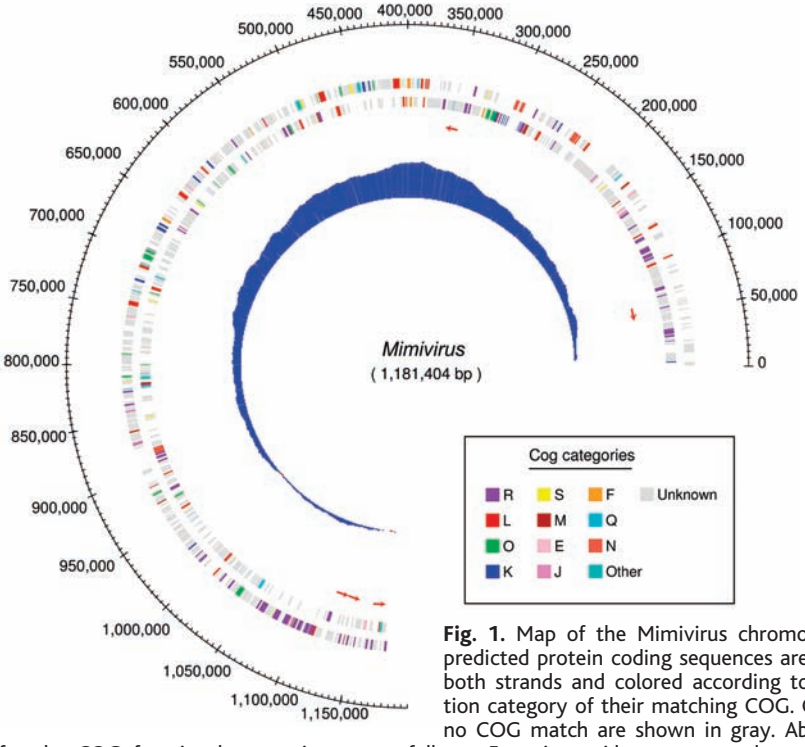


Fig. 1. Map of the Mimivirus chromosome. The predicted protein coding sequences are shown on both strands and colored according to the function category of their matching COG. Genes with no COG match are shown in gray. Abbreviations for the COG functional categories are as follows: E, amino acid transport and metabolism; F, nucleotide transport and metabolism; J, translation; K, transcription; L, replication, recombination, and repair; M, cell wall/membrane biogenesis; N, cell motility; O, posttranslational modification, protein turnover, and chaperones; Q, secondary metabolites biosynthesis, transport, and catabolism; R, general function prediction only; S, function unknown. Small red arrows indicate the location and orientation of tRNAs. The A+C excess profile is shown on the innermost circle, exhibiting a peak around position 380,000 (*2*) (fig. S1).

program (*7*). We found that 194 Mimivirus ORFs exhibited significant matches with 108 distinct COG families (table S3). This is more than twice the number of COGs represented in PBCV-1 virus (46 ORFs matching with 41 COGs). Compared with other NCLDVs, Mimivirus COG profile exhibits a significant overrepresentation in the functional categories of translation (COG category J), posttranslation modifications (COG category O), and amino acid transport and metabolism (COG category E) (X^2 test: $P < 0.001$, $P = 0.006$, and $P = 0.08$, respectively) (*2*) (table S3).

Features in the Mimivirus Genome Unique Among dsDNA Viruses

The detailed analysis of Mimivirus genome (*2*) revealed a number of unique features, including many genes never before identified in a viral genome. Until now, some of these genes were thought to be the trademark of cellular organisms. These previously unknown and unique genes are listed in Table 2. They can be classified in four generic functional categories: protein translation, DNA repair enzymes, chaperones, and new enzymatic pathways. In addition, Mimivirus is the sole virus and one of the rare microorganisms that simultaneously possesses type IA, type IB, and type II topoisomerases.

Protein translation–related genes. The inability to perform protein synthesis independently from their host is one of the main characteristics distinguishing viruses from cellular ("living") organisms. However, tRNA-like genes are found in isolated dsDNA viruses species such as bacteriophage T4 (*8*) and BxZ1 (*9*), herpes virus 4 (*10*), and chlorella viruses (*11*). The chlorella viruses are also the first ones found to encode a translation elongation factor (EF-3) (*12*). The genome analysis of Mimivirus now greatly expands the known repertoire of viral genes related to protein translation. In addition to six tRNA-like genes [three Leu (two TTAs, and one TTG), Trp (TGG), Cys (TGC), and His (CAC)], the Mimivirus genome exhibits homologs to 10 proteins with functions central to protein translation: four aminoacyl-tRNA synthetases (aaRSs), translation initiation factor 4E (e.g., mRNA cap–binding), translation factor eF-TU [guanosine 5′-triphosphate (GTP)–binding translocation factor], translation initiation factor SUI1, translation initiation factor IF-4A (a helicase), and peptide chain release factor eRF1. In addition, the Mimivirus genome encodes the first identified viral homolog of a tRNA modifying enzyme (tRNA (Uracil-5-)–methyltransferase). All of these ORFs have significant sequence similarity with their eukaryotic homologs and exhibit all the domains and specific signatures expected from functional representatives of these various gene families. Preliminary functional characterizations have been obtained for several of these genes. For instance, we produced Mimivirus tyrosyl-tRNA synthetase in *Escherichia coli*, purified it, and measured its enzymatic activity (*2*) (fig. S2). Crystals of the protein have been obtained and its three-dimensional (3D) structure has been determined (*13*). In addition, mRNAs encoding Mimivirus tyrosyl-, cysteinyl-, and arginyl-tRNA synthetases are found associated with purified virus particles (*2*) (table S4), suggesting that they are involved in infection.

New DNA repair enzymes. Genomes are subject to damage by chemical mutagens (e.g., free radicals alkylating agents), ultraviolet (UV) light, or ionizing radiations. Different repair pathways have evolved to prevent the lethal accumulation of the various types of DNA errors. They usually correspond to well-conserved protein families found in the three domains of life (Archaea, Eubacteria, and Eukaria) but to a much lesser extent in viruses. The analysis of the Mimivirus genome revealed several types of DNA repair enzyme homologs, including four never before reported in dsDNA viruses. For instance, we identified two genes (L315 and L720) encoding putative formamidopyrimidine-DNA glycosylases, which serve to locate and excise oxidized purines. The Mimivirus genome also exhibits a UV-damage endonuclease (UvdE)

Table 1. NCLDV core genes (classes I, II, and III) identified in Mimivirus. Black squares, best matching homologs; X, significant homolog detected in all available genomes; x, not in all in available genomes; sub., subunit.

ORF no.	*Phycodnaviridae*	*Poxviridae*	*Irido viridae*	*Asfar viridae*	Gene group	Definition/putative function (*5*)
L206	X	X	■	X	I	Helicase III / VV D5-type ATPase
R322	■	X	X	X	I	DNA polymerase (B family)
L437	X	X	■	X	I	VV A32 virion packaging ATPase
L396	■	X	x	X	I	VV A18 helicase
L425	■	X	X	X	I	Capsid protein D13L (4 paralogs)
R596	■	X	X	X	I	Thiol oxidoreductase (e.g., E10R)
R350	X	■	X	X	I	VV D6R helicase, +1paralog
R400	■	X	X	X	I	S/T protein kinase (e.g., F10L)
R450	■	X	X	X	I	Transcription factor (e.g., A1L)
R339	■x	X	X	X	II	TFII-like transcription factor
L524	x	X	■X	X	II	MuT-like NTP pyrophosphohydrolase
L323	x	X	■X	X	II	Myristoylated virion protein A
R493	■X	x	X	X	II	PCNA + 1 paralog
R313	X	■x	X	X	II	Ribonucleotide reductase, large sub.
L312	X	■x	X	X	II	Ribonucleotide reductase, small sub.
Not found	x	x	X	X	II	Thymidylate kinase
Not found	x	X	X	X	II	dUTPase
R429	■	–	X	X	III	PBCV1-A494R-like (9 paralogs)
L37	X	■	X	X	III	BroA, KilA-N term
R382	X	X	–	■	III	mRNA–capping enzyme
L244	–	X	■	X	III	RNA polymerase subunit 2 (Rbp2)
R501	–	X	■	X	III	RNA polymerase largest sub. (Rpb1)
R195	■	X	X	–	III	Glutaredoxin (e.g., ESV128)
R622	X	■	X	–	III	Dual spec. S/Y phosphatase
R311	–	x	X	X	III	BIR domain (e.g., CIV193R)
L65	–	■X	X	X	III	Virion-associated membrane protein
R480	■	–	X	X	III	Topoisomerase II
L364	X	■	X	–	III	SW1/SNF2 helicase (e.g., MSV224)
Not found	x	X	X	–	III	RuvC-like HJR (e.g., A22R)
Not found	x	x	–	X	III	ATP-dependent DNA ligase (e.g., A50R)
Not found	–	x	X	X	III	RNA polymerase subunit 10

homolog (L687). Although this is the first report of such an enzyme in a dsDNA virus, we identified an isolated UvdE homolog among the "hypothetical" proteins of the recently sequenced *Aeromonas hydrophila* phage Aeh1 (ORF111c, GenBank accession code: AAQ17773). The major mutagenic effect of methylating agents in DNA is the formation of O^6-alkylguanine. The corresponding repair is performed by a DNA-[protein]-cysteine *S*-methyltransferase. The Mimivirus genome encodes the first viral 6-*O*-methylguanine-DNA methyltransferase (R693). In addition, Mimivirus R406 ORF is strongly homologous to a number of bacterial genes annotated as belonging to the same alkylated DNA repair pathways. Finally, ORF L359 was found to clearly belong to the MutS protein family, which is involved in DNA mismatch repair and recombination. Again, this is the first DNA repair enzyme of this family described in a dsDNA virus. Aside from the above DNA repair system components, which have never before been reported in dsDNA virus, Mimivirus ORF L386 and R555 encode homologs to the rad2 and rad50 yeast genes, respectively, both central to the repair of UV-induced DNA damage. Homologs for these genes are also found in Iridoviruses. Overall, Mimivirus appears uniquely well equipped to repair DNA mismatch and damages caused by oxidation, alkylating agent, or UV light.

Topoisomerases. DNA topoisomerases are the enzymes in charge of solving the topological (entanglement) problems associated with DNA replication, transcription, recombination, and chromatin remodeling (*14*). Type I topoisomerases (ATP independent) work by passing one strand of the DNA through a break in the opposite strand. Type II topoisomerases are adenosine triphosphatases (ATPases) and work by introducing a double-stranded gap. Topoisomerases of various types are involved in relaxing or introducing DNA supercoils. With the notable exception of *Poxviridae*, many dsDNA viruses (including NCLDVs and phages) encode their own type IIA topoisomerase. Accordingly, Mimivirus exhibits a large ORF (>1263 amino acids, R480) 41% identical to PBCV-1 topoisomerase IIA amino acid sequence. Its best database match overall is with a homologous protein in the small eukaryote *Encephalitozoon cuniculi* (42% identical). More surprisingly, Mimivirus is the first dsDNA virus found to also encode a *Poxviridae*-like topoisomerase (topoisomerase IB). Mimivirus ORF R194 is 27% identical to *Amsacta moorei* entomopoxvirus topoisomerase IB (AMV052) and 25% identical to the well-studied vaccinia virus topoisomerase (H6R). In addition, to encode both type IIA and type IB topoisomerases, Mimivirus exhibits the first type IA topoisomerase reported in a virus (*14*). The ORF L221 best overall database match (37%) is with its homolog in *Bacteroides thetaiotaomicron* (a Gram-negative anaerobe colonizing the human colon) within a well-defined subgroup of well-conserved type IA eubacterial topoisomerases, the prototype of which is *E. coli* Omega untwisting enzyme. Among all available genome sequences, only a small number of microorganisms simultaneously exhibit topoisomerases of type IA, IB, and IIA. They include yeast, *Deinococcus radiodurans*, and various environmental bacteria such as *Pseudomonas* sp., *Agrobacterium tumefaciens*, and *Sinorhizobium meliloti*.

Protein folding. The folding of many proteins, in particular those involved in large molecular assemblies, is guided toward their native structures by different families of protein chaperones. The Mimivirus genome uniquely exhibits two ORFs entirely and highly homologous to chaperones of the HSP70 (DnaK) family. ORF L254 is 42% identical to DnaK protein 2 of *Thermosynechococcus elongates*, and ORF L393 is 59% identical to bovine heat-shock 70-kD protein 1A. In addition, the Mimivirus genome exhibits three ORFs (R260, R266, and R445) with clear DnaJ domain signatures. Proteins containing a DnaJ domain are known to associate with proteins of the HSP70 family. The above Mimivirus ORFs might thus encode a set of proteins interacting to form a specific viral chaperone

Table 2. Major new features identified in Mimivirus genome. dTDP, 3′-deoxy-thymidine-5′diphosphate; ADP, adenosine 5′-diphosphate.

ORF no.	Definition/putative function	Comment
R663	Arginyl-tRNA synthetase	Translation
L124	Tyrosyl-tRNA synthetase	Translation
L164	Cysteinyl-tRNA synthetase	Translation
R639	Methyonyl tRNA synthetase	Translation
R726	Peptide chain release factor eRF1	Translation
R624	GTP-binding elongation factor eF-Tu	Translation
R464	Translation initiation factor SUI1	Translation
L496	Translation initiation factor 4E (mRNA cap binding)	Translation
R405	tRNA (Uracil-5-)-methyltransferase	tRNA modification
L359	DNA mismatch repair ATPase MutS	DNA repair
R693	Methylated-DNA-protein-cysteine methyltransferase	DNA repair
R406	Alkylated DNA repair	DNA repair
L687	Endonuclease for the repair of UV-irradiated DNA	DNA repair
L315 L720	Hydrolysis of DNA containing ring-opened N7-methylguanine	DNA repair
R194 R480 L221	Topoisomerase I pox-like, topoisomerase II, topoisomerase I bacterial type	DNA accessibility
L254 L393	Heat shock 70-kD	Chaperonin
L605	Peptidylprolyl isomerase	Chaperonin
L251	Lon domain protease	Chaperonin
R418	NDK synthesis of nucleoside triphosphates	Metabolism
R475	Asparagine synthase (glutamine hydrolyzing)	Metabolism
R565	Glutamine synthetase (Glutamate-amonia ligase)	Metabolism
L716	Glutamine amidotransferase domain	Metabolism
R689	*N*-acetylglucosamine-1-phosphate, uridyltransferase	Polysaccharide synthesis
L136	Sugar transaminase, dTDP-4-amino-4,6-dideoxyglucose biosynthesis	ExoPolysaccharide synthesis
L780	dTDP-4-dehydrorhamnose reductase	ExoPolysaccharide synthesis
L612	Mannose-6P isomerase	Glycosylation
L230	Procollagen-lysine,2-oxoglutarate 5-dioxygenase	Glycosylation, capsid structure
L543	ADP-ribosyltransferase (DraT)	?
L906	Cholinesterase	Host infection?
L808	Lanosterol 14-alpha-demethylase	Host infection?
R807	7-dehydrocholesterol reductase	Host infection?
R322	Intein insertion	In DNA polymerase B

system, possibly required for the productive assembly of its huge capsid.

In addition to its gene equipment related to protein folding, Mimivirus is the first to encode a homolog to the lon *E. coli* heat-shock protein, an ATP-dependent protease thought to dispose of unfolded polypeptides. Mimivirus also exhibits components of the ubiquitin-dependent protein degradation pathway, already described in other NCLDVs. Finally, the Mimivirus genome encodes a putative peptidyl-prolyl cis-trans isomerase of the Cyclophilin family (ORF L605). This type of enzyme, seen here in a virus for the first time, accelerates protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds. Again, this new virally encoded function might be required for the Mimivirus capsid to be assembled within physiological time limits.

New metabolic pathways. The genome analyses of large Phycodnaviruses and other NCLDVs already contributed the notion that large viruses possess significant metabolic pathways in addition to the minimal infection, replication, transcription, and virion packaging systems. PBCV-1, for instance, exhibits enzymes for the synthesis of homospermidine, hyaluronan, guanosine diphosphate (GDP)–fucose, and many other sugar-, lipid-, and amino acid–related manipulations (*15*). With its larger genome, Mimivirus builds on this established trend by exhibiting previously described as well as new virally encoded biosynthetic capabilities.

For instance, Mimivirus genome encodes homologs to many enzymes related to glutamine metabolism: asparagine synthase (glutamine hydrolyzing) (ORF R475), glutamine synthase (ORF R565), and guanosine 5′-monophosphate synthase (glutamine hydrolyzing) (ORF L716). All are identified in a dsDNA virus for the first time. In addition, Mimivirus exhibits a glutamine: fructose-6-P aminotransferase (i.e., glucosamine synthase) as previously described in PBCV-1. Mimivirus can proceed further along this pathway with the use of its own encoded *N*-acetylglucosamine-1-phosphate uridyltransferase (the well-studied GlmU enzyme) (ORF R689) to synthesize uridine 5′-diphosphate–*N*-acetyl-glucosamine. This metabolite is central to the biosynthesis of all types of polysaccharides in both eukaryotic and prokaryotic systems. The Mimivirus genome encodes six glycosyltranferases: three from family 2, and one each from families 8, 10, and 25.

Glycosyltransferases form a complex group of enzymes involved in the biosynthesis of disaccharides, oligosaccharides, and polysaccharides that are involved in the posttranslational modification of proteins (N- and O-glycosylation), and the synthesis of lipopolysaccharides included in high–molecular weight cross-linked periplasmic or capsular material. Among other NCLDVs, PBCV-1 has been well studied in that respect and shown to encode an atypical N-glycosylation pathway and hyaluronan biosynthesis (*15*). Other chloroviruses promote the synthesis of chitin (*16*). Preliminary proteomic studies of Mimivirus particles (see below) indicate that several proteins are glycosylated, including the predicted major capsid protein. In addition, Mimivirus particles are positive upon standard Gram staining (*1*), suggesting the presence of a reticulated polysaccharide at their surface. It is likely that some of the Mimivirus glycosyltranferases are involved in its synthesis. For instance, Mimivirus encodes (L136) a homolog to perosamine synthetase. Such an enzyme catalyzes the conversion of GDP-4-keto-6-deoxymannose to 4-NH_2-4,6-dideoxymannose (perosamine), which is found in the O-antigen moiety of the lipopolysaccharide of various bacteria. Another Mimivirus ORF (L230) is homologous to procollagen-lysine, 2-oxoglutarate 5-dioxygenase. This enzyme catalyzes the formation of hydroxylysine in collagens and other proteins with collagen-like amino acid sequences by the hydroxylation of lysine residues in X-Lys-Gly sequences. These hydroxyl groups then serve as sites of attachment for carbohydrate units and are also essential for the stability of the intermolecular collagen cross-links. Given that Mimivirus also contains a large number of ORFs exhibiting the characteristic collagen triple-helix repeat, it is tempting to speculate that the hairy-like appearance of the virion (*1*) might be due to a layer of cross-linked glycosylated collagen-like fibrils.

Among other enzymes never yet reported in a virus, Mimivirus includes a NDK [Enzyme Classification (EC): 2.7.4.6] (ORF R418). NDK catalyzes the synthesis of nucleoside triphosphates (NTPs) other than ATP. This enzyme may help circumvent a limited supply of NTPs for nucleic acid synthesis, UTP for polysaccharide synthesis, and GTP for protein elongation.

Finally, Mimivirus is also encoding homologs to three lipid-manipulating enzymes: cholinesterase (L906), lanosterol 14-alpha-demethylase (L808), and 7-dehydrocholesterol reductase (R807), the physiological roles of which remain to be determined but possibly include the disruption of the host membrane.

Intein and introns. Inteins are protein-splicing domains encoded by mobile intervening sequences (IVSs) (*17*). They self-catalyze their excision from the host protein, ligating their former flanks by a peptide bond. They have been found in all domains of life (Eukaria, Archaea, and Eubacteria), but their distribution is highly sporadic. Only a few instances of viral inteins have been described, in *Bacillus subtilis* bacteriophages (*18*) and in the ribonucleotide reductase alpha subunit of *Chilo* iridescent virus (CIV) (*19*). Mimivirus is then the second eukaryotic dsDNA virus exhibiting an intein (*2*). In contrast with the one described for CIV (lacking a C-terminal Asn), Mimivirus intein is canonical and exhibits valid amino acids at all essential positions, as well as the dodecapeptide homing endonuclease motif (*20*). For reasons not yet understood, inteins are most often found associated with essential enzymes of the DNA metabolism. Inserted within DNA polymerase B, Mimivirus intein is no exception to this rule.

Self-splicing type I introns are a different type of mobile IVS, self-excising at the mRNA level. They are rare in viruses and mostly found in phages. One type IB intron has been identified in several chlorella virus species (*15*). Mimivirus exhibits four instances of self-excising intron (*2*), all in RNA polymerase genes: One in the largest and three in the second-largest subunit.

Gene families or protein domains expanded in Mimivirus. The ankyrin-repeat signature is the most frequent motif, found in more than 30 distinct ORFs. This motif, about 33 amino acids long, is one of the most common protein-protein interaction motifs. It has been found in proteins with a wide diversity of functions. Another protein interaction domain, defined by the BTB signature, is found in 20 ORFs. This domain mostly mediates homomeric dimerization. It is found in proteins that contain the KELCH motif such as Kelch and a family of pox virus proteins. We identified 14 different ORFs exhibiting the protein kinase motif (PFAM) signature (*21*) of the catalytic domain of eukaryotic protein kinases ($P < 0.05$). Four of them resemble known cell division–related kinases.

The collagen triple-helix motif is another frequently represented motif, found in eight ORFs. This motif is characteristic of extracellular structural proteins involved in matrix formation and/or adhesion processes. Like other collagens, the product of these collagen-like ORFs might be posttranslationally modified by the procollagen-lysine, 2-oxoglutarate 5-dioxygenase homolog uniquely found in Mimivirus genome. Mimivirus also contains eight ORFs with significant similarity to helicases. Finally, Mimivirus exhibits eight ORFs containing a specific glucose-methanol-choline (GMC) oxidoreductase motif. The role of these flavin adenine dinucleotide flavoproteins is unknown.

Phylogeny

Relationship to other NCLDVs. Our preliminary study based on the protein sequences of ribonucleotide reductase small and large subunits and topoisomerase II (*1*) suggested an independent branching of Mimivirus in the phylogenetic tree of NCLDVs (*1*). This analysis was refined by using the concatenated sequences of the eight ''class I'' genes conserved in Mimivirus and all other

NCLDVs. The resulting phylogenetic tree again suggested that Mimivirus defines an independent lineage of NCLDVs (Fig. 2) roughly equidistant from known Phycodnaviruses and Iridoviruses.

Relationship to the three domains of life. There are 63 COGs common to all known unicellular genomes from the three domains of life: Eukarya, Eubacteria, and Archaea. Seven of them are now identified in the genome of Mimivirus: three aminoacyl-tRNA synthetases [ArgRS (COG0018), MetRS (COG0143), and TyrRS (COG0162)], the beta (COG0085) and beta′ (COG0086) subunits of RNA polymerase, the sliding clamp subunit of DNA polymerase [three proliferating cell nuclear antigen (PCNA) paralogs; COG0592], and a 5′-3′ exonuclease (COG0258). The unrooted phylogenetic tree built from the concatenated sequences of those proteins (*2*) is shown in Fig. 3. Mimivirus branches out near the origin of the Eukaryota domains. This is supported with a high bootstrap value and the Shimodaira-Hasegawa statistical test (*2*). The tree topology is also invariant to a variety of methodological changes (*2*) [figs. S3 to S6 and supporting online material (SOM) text]. Consistently, scatter plots for the best BLAST scores against the three domains of life indicate that most Mimivirus ORFs exhibit higher sequence similarities to eukaryotic sequences than to prokaryotic sequences, and are equidistant from the four main eukaryotic kingdoms: Protista, Animalia, Plantae, and Fungi (*2*) (fig. S7). However, strictly speaking, the tree shown in Fig. 3 can be rooted on any of the deepest branches, including the branch separating Mimivirus from eukaryotes, making its specific affinity with Eukaryota still uncertain.

Genome Complexity: Mimivirus Versus Parasitic Cellular Organisms

The number of Mimivirus COGs was compared to the numbers found for representatives of the three domains of life with the smallest known genomes: *Nanoarchaeum equitans* (490 kb), *Mycoplasma genitalium* (580 kb), and *Encephalitozoon cuniculi* (2.498 kb) (Fig. 4). Despite its comparable genome size, Mimivirus exhibits fewer identified COGs. However, there was no specific category in which it was significantly underrepresented, except for the translation category ($P < 0.01$). By this standard, the absence of a functional protein-translation apparatus is what most distinguishes Mimivirus from its parasitic cellular counterparts.

Preliminary Analysis of Mimivirus Particles

Detection of viral RNAs. Large viruses such as those of the *Herpesviridae* family, incorporate viral transcripts during the particle assembly process (*22*). We thus investigated whether viral RNAs could be found associated with ribonuclease (RNase)–treated Mimivirus particles with the use of reverse transcription polymerase chain reaction and virus-specific primers targeting several genes (*2*). Positive results were obtained for three aminoacyl tRNA synthetases (TyrRS, CysRS, and ArgRS), DNA polymerase, transcription factor TFIIB, and the predicted major capsid protein gene (L425) (*2*) (table S4).

Virion proteomics. Constituent proteins of Mimivirus particles were extracted and analyzed. In a preliminary set of experiments, 2D gel electrophoresis resolved 438 spots, many of them visibly corresponding to multiple isoforms of the same gene product (such as glycosylation and phosphorylation) (*2*) (fig. S8). The most abundant of the best-resolved spots were eluted and characterized by mass spectrometry (Maldi-ToF and ion trap). Six predicted ORF products corresponding to proteins with homologs of known functions were unambiguously identified. As expected, they include the major capsid (L425) and core (L410) proteins but also an mRNA-capping enzyme (R382), thioredoxin (R548), and glutaredoxin (R195), and a GMC-type oxidoreductase (R135).

Virion resistance to adverse conditions. Mimivirus particles remained infectious during 1 year when kept at 4°C, 25°C, and 32°C in Page's amoeba saline (PAS) buffer. Incubation of a suspension of 10^9 particles in PAS buffer at 55°C from 15 to 90 min reduced its titer by 100. By comparison, no viable *E. coli* are retrieved when submitted to the same treatment. No diminution in Mimivirus titer was observed after 48 hours desiccation. Mimivirus particles are thus quite resistant to adverse conditions. However, despite its many predicted DNA repair genes, Mimivirus is quickly killed by 35 kilograys of irradiation with gamma rays or exposure for 15 min (30 W, 20 cm) to UV light (*2*).

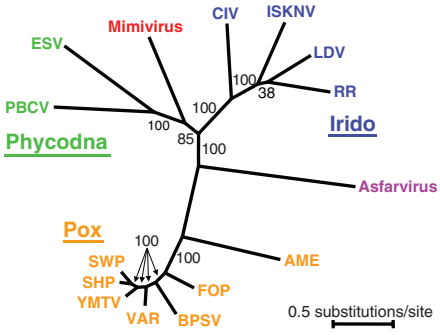


Fig. 2. Phylogenetic position of Mimivirus among established NCLDV families. Viral species representing the diverse families of NCLDV are included as follows: Mimivirus, *Phycodnaviridae* (PBCV and ESV), *Iridoviridae* [CIV, *Regina* ranavirus (RR), lymphocystis disease virus type 1 (LDV), and infectious spleen and kidney necrosis virus (ISKNV)], *Asfarviridae* (African swine fever virus), and *Poxviridae* [*Amsacta moorei* entomopoxvirus (AME), variola virus (VAR), fowlpox virus (FOP), bovine papular stomatitis virus (BPSV), Yaba monkey tumor virus (YMTV), sheeppox virus (SHP), and swinepox virus (SWP)]. Fully sequenced viral genomes were analyzed to ensure the proper assessment of orthologous genes. This tree was built with the use of maximum likelihood and based on the concatenated sequences of eight conserved proteins (NCLDV class I genes): vaccina virus (VV) D5-type ATPase, DNA polymerase family B, VV A32 virion packaging ATPase, capsid protein, thiol oxidoreductase, VV D6R helicase, serine/threonine protein kinase, and A1L transcription factor. One of the class I genes (VV A18 helicase) was absent in LDV and was not included. The alignment contains 1660 sites without insertions and deletions. A neighbor joining tree and a maximum parsimony tree exhibited similar topologies (*2*). Bootstrap percentages are shown along the branches.

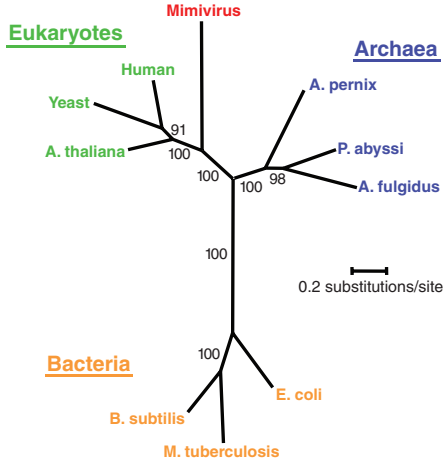


Fig. 3. A phylogenetic tree of species from the three domains of life (Eukaryota, Eubacteria, and Archaea) and Mimivirus. The tree was inferred with the use of a maximum likelihood method based on the concatenated sequences of seven universally conserved protein sequences: arginyl-tRNA synthetase (COG0018), methionyl-tRNA synthetase (COG0143), tyrosyl-tRNA synthetase (COG0162), RNA polymerase II largest subunit (COG0086), RNA polymerase II second largest subunit (COG0085), PCNA (COG0592), and 5′-3′ exonuclease (COG0258). The alignment contains 3164 sites without insertions and deletions. Bootstrap percentages are shown along the branches. Similar trees were obtained with the use of a variety of other approaches (SOM text).

Discussion

A common feature to all known viruses is their total dependency on the host translation machinery for protein synthesis. Surprisingly, the Mimivirus genome sequence now reveals genes relevant to all key steps of mRNA translation: tRNA and tRNA charging, initiation, elongation, and termination, with the exception of ribosome components themselves. Two main evolutionary scenarios may account for the presence of this partial complement of translation-related genes in Mimivirus. On one hand, they could

be the relics of a more complete ancestral protein-translation apparatus, gradually lost through a genome reduction process similar to the one governing the evolution of intracellular bacteria (*23*). On the other hand, these genes could have been individually acquired from cellular organisms and used to control the host translation apparatus in favor of Mimivirus mRNAs. The evidence that our phylogenetic analysis did not support a recent acquisition of these genes, together with the low probability that these genes were acquired independently, is in favor of the loss rather than the gain scenario. By extrapolating this model, we could speculate that the Mimivirus lineage originated from a more complex ancestor possibly exhibiting an even more complete protein-translation machinery.

By its particle size, and now by its genome complexity, Mimivirus significantly challenges our vision of viruses. Lwoff (*24*) proposed that viruses should have at least one dimension lower than 200 nm and speculated that viruses may possess only one type of nucleic acid. Both criteria are invalidated by Mimivirus. Lwoff also pointed out the lack of enzymes generating energy from substrates. This criteria is still valid, because very few genes of this category were detected in Mimivirus. Other criteria such as the strictly intracellular character and the inability to grow or undergo binary fission have not yet been challenged. By these three last criteria, Mimivirus remains a regular virus. However, by the unprecedented number of enzymes and putative metabolic pathways encoded by its 1.2-Mb genome, Mimivirus blurs the established frontier between viruses and the parasitic cellular organisms with small defective genomes such as *Rickettsia prowazekii* (*25*), Buchnera (*26*), *Nanoarchaeum* (*27*), Mycoplasma (*28*), and *Tropheryma whipplei* (*29*). As of today, the genome of Mimivirus is larger than the published genomes of 20 cellular organisms from two domains of life (e.g., Archaea and Eubacteria) and five main bacterial divisions: Proteobacteria, Firmicutes, Actinobacteria, Chlamydiae, and Spirochaetes. The presence versus absence of ribosomes remains, at the moment, a key property distinguishing these minimal cellular organisms from large DNA viruses.

Several independent studies have led to the hypothesis that DNA viruses may have a common origin, and a common ancestor, originating before the emergence of the three domains of life (*30*). Given the inherent uncertainty of phylogenetic reconstruction dating back 3 billion years ago, our results (Fig. 3) are consistent with the hypotheses that a lineage of large DNA viruses could have emerged before the individualization of cellular organisms from the three domains of life (*31*) or from an ancestor distinct of these three domains (*32*). The topology of this new ''tree of life'' is also consistent with the hypothesis that ancestral DNA viruses were involved in the emergence of Eukaryotes (*33–37*).

The serendipitous discovery of Mimivirus from samples initially thought to contain a new type of intracellular Gram-positive bacterium allowed the characterization of the largest virus so far. The sequencing of its 1.2-Mb genome revealed a wealth of genes encoding functions never yet encountered in viruses, probably due to its unprecedented size. The numerous new genes related to the protein-translation apparatus challenge the established vision of viruses. Using these first viral representatives of universally conserved gene families, we could now build a tentative tree of life, within which Mimivirus appears to define a new branch distinct from the three other domains. We believe that our work should prompt the search for more giant viruses, the genome analysis of which could shed additional light on the origin of DNA viruses and their role in the evolution of cellular organisms.

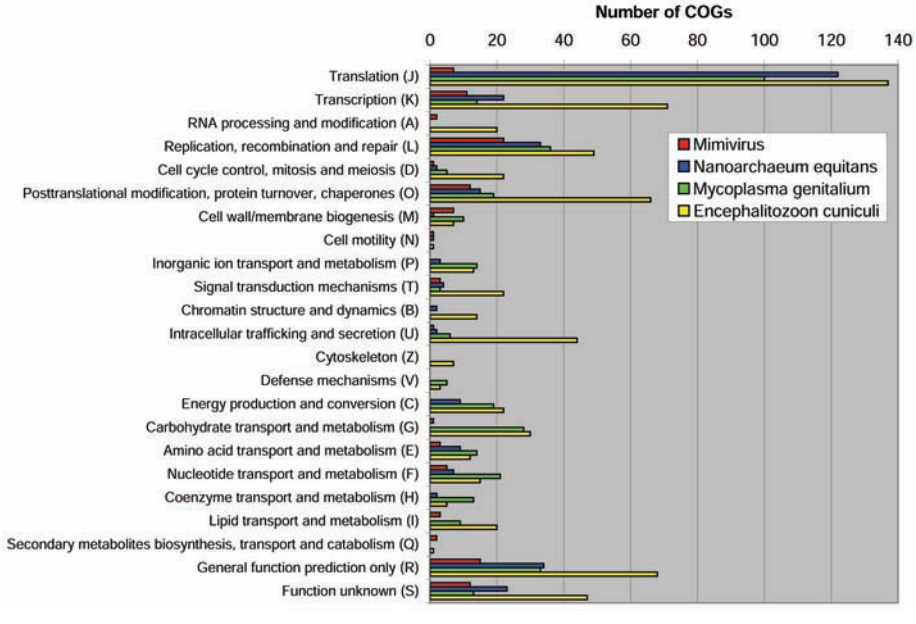


Fig. 4. Distribution of COG homologs in Mimivirus compared with the cellular organisms of the three domains of life with the smallest known genomes.

References and Notes

1. B. La Scola *et al.*, *Science* **299**, 2033 (2003).
2. Materials and methods are available as supporting material on *Science* Online.
3. N. Nandhagopal *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14758 (2002).
4. C. L. Afonso *et al.*, *J. Virol.* **73**, 533 (1999).
5. L. M. Iyer, L. Aravind, E. V. Koonin, *J. Virol.* **75**, 11720 (2001).
6. R. L. Tatusov *et al.*, *BMC Bioinformat.* **4**, 41 (2003).
7. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
8. E. S. Miller *et al.*, *Microbiol. Mol. Biol. Rev.* **67**, 86 (2003).
9. M. L. Pedulla *et al.*, *Cell* **113**, 171 (2003).
10. H. W. Virgin *et al.*, *J. Virol.* **71**, 5894 (1997).
11. J. L. Van Etten, R. H. Meints, *Annu. Rev. Microbiol.* **53**, 447 (1999).
12. T. Yamada, T. Fukuda, K. Tamura, S. Furukawa, P. Songsri, *Virology* **197**, 742 (1993).
13. C. Abergel *et al.*, in preparation.
14. J. J. Champoux, *Annu. Rev. Biochem.* **70**, 369 (2001).
15. J. L. Van Etten, *Annu. Rev. Genet.* **37**, 153 (2003).
16. T. Kawasaki *et al.*, *Virology* **302**, 123 (2002).
17. S. Pietrokovski, *Trends Genet.* **17**, 465 (2001).
18. V. Lazarevic *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1692 (1998).
19. S. Pietrokovski, *Curr. Biol.* **8**, R634 (1998).
20. H. Ogata *et al.*, in preparation.
21. A. Bateman *et al.*, *Nucleic Acids Res.* **30**, 276 (2002).
22. W. A. Bresnahan, T. Shenk, *Science* **288**, 2373 (2000).
23. N. A. Moran, *Cell* **108**, 583 (2002).
24. A. Lwoff, *J. Gen. Microbiol.* **17**, 239 (1957).
25. S. G. E. Andersson *et al.*, *Nature* **396**, 133 (1998).
26. I. Tamas I *et al.*, *Science* **296**, 2376 (2002).
27. E. Waters *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12984 (2003).
28. C. M. Fraser *et al.*, *Science* **270**, 397 (1995).
29. D. Raoult *et al.*, *Genome Res.* **13**, 1800 (2003).
30. W. Zillig *et al.*, *Extremophiles* **2**, 131 (1998).
31. C. Woese, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6854 (1998).
32. D. P. Mindell, L. P. Villarreal, *Science* **302**, 1677 (2003).
33. L. P. Villarreal, V. R. DeFilippis, *J. Virol.* **74**, 7079 (2000).
34. P. Forterre, *C. R. Acad. Sci. III* **324**, 1067 (2001).
35. M. Takemura, *J. Mol. Evol.* **52**, 419 (2001).
36. P. J. L. Bell, *J. Mol. Evol.* **53**, 251 (2001).
37. E. Pennisi, *Science* **305**, 766 (2004).

38. We thank C. Fraisier, A. Honstettre, V. Arondel, N. Aldrovandi, S. Chenivesse, and D. Moinier for technical help; and M. Drancourt and K. Suhre for helpful discussions. We acknowledge funding from Marseille-Nice Genopole. The IGS laboratory is partially supported by Aventis Pharma. The Mimivirus genome sequence has been deposited to GenBank under accession number AY653733. Additional data are accessible at www.giantvirus.org.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1101485/DC1
Materials and Methods
SOM Text
Figs. S1 to S8
Tables S1 to S4
References

15 June 2004; accepted 22 September 2004
Published online 14 October 2004;
10.1126/science.1101485
Include this information when citing this paper.

Magnetic Response of Metamaterials at 100 Terahertz

Stefan Linden,[1] Christian Enkrich,[2] Martin Wegener,[1,2] Jiangfeng Zhou,[3] Thomas Koschny,[3,4] Costas M. Soukoulis[3,4]*

An array of single nonmagnetic metallic split rings can be used to implement a magnetic resonance, which arises from an inductor-capacitor circuit (LC) resonance, at 100-terahertz frequency. The excitation of the LC resonance in the normal-incidence geometry used in our experiments occurs through the coupling of the electric field of the incident light to the capacitance. The measured optical spectra of the nanofabricated gold structures come very close to the theoretical expectations. Additional numerical simulations show that our structures exhibit a frequency range with negative permeability for a beam configuration in which the magnetic field couples to the LC resonance. Together with an electric response that has negative permittivity, this can lead to materials with a negative index of refraction.

Our ability to tailor the potential of electrons on the scale of their de Broglie wavelength has opened the door to new frontiers in nanoelectronics. Similarly, our ability to tailor the index of refraction of materials is the route to new avenues in nanophotonics. Usually, the index of refraction n determines the factor to which the propagation of light in a medium is slower than in vacuum. Hence, one traditionally expects n to be a positive number. Most of the time, n is larger than unity. With this in mind, the recently realized metamaterials (*1*), which have a negative index of refraction (*2*–*6*), came as a surprise to many. The negative n in these so-called left-handed materials (*2*) can lead to new physical phenomena and potential applications, such as "perfect lenses" (*3*).

Under conditions of negative refraction, a light wave impinging from vacuum or air onto the metamaterial's surface under an angle with respect to the surface normal is refracted toward the "wrong" side of the normal (*1*). A negative index of refraction, n, in Snell's law indeed reproduces this unusual behavior. Mathematically, the square of the index of refraction, $n^2 = \varepsilon\mu$, is given by the product of the electric permittivity ε and the magnetic permeability μ of the medium. If both permittivity and permeability are negative, the resulting refractive index is negative as well. A negative permittivity is not unusual and occurs in any metal from zero frequency to the plasma frequency; however, a large magnetic response, in general, and a negative permeability at optical frequencies, in particular, do not occur in natural materials.

It is of particular importance to terahertz optics and their applications (*7*) to achieve magnetic resonant response at terahertz and higher frequencies. In metamaterials, this crucial aspect is achieved by mimicking an LC oscillator of eigenfrequency ω_{LC} with $\omega_{LC} = (LC)^{-1/2}$, consisting of a magnetic coil with inductance L and a capacitor with capacitance C (Fig. 1). The incident light can couple to the LC resonance (*8*), if at least one of the following conditions is fulfilled (Fig. 1): (i) The electric field vector $\boldsymbol{E}$ of the incident light has a component normal to the plates of the capacitor, or (ii) the magnetic field vector $\boldsymbol{H}$ of the incident light has a component normal to the plane of the coil. If condition (ii) is fulfilled, the current in the coil, analogous to an atomic orbital current, leads to a magnetic field that counteracts the driving magnetic field, which can lead to a negative permeability. Such metamaterials were first realized at frequencies around 10 GHz (3-cm wavelengths) (*1*, *9*) and could be fabricated on stacked electronic circuit boards.

It was believed that corresponding materials at optical frequencies, which are more than four orders of magnitude higher (a few hundred terahertz), were out of reach because of Ohmic losses. However, microstructures with magnetic resonance frequencies that were two orders of magnitude larger at about 1 THz (300-μm wavelength) were recently fabricated by microlithography (*10*). Using nanofabrication techniques, we increased the LC-resonance frequency by yet another two orders of magnitude to about 100 THz (3-μm wavelength), bringing optical frequencies into reach for obtaining negative index of refraction. Both transmission and reflection were measured. The measured transmission spectra gave more than 90% transmission, so the losses are relatively low for these very thin metallic structures. The comparison of the measured optical spectra with theory shows very good agreement.

Our design of the structures closely follows a recent theoretical suggestion (*8*), which uses single split-ring resonators (SRRs) rather than double SRRs (*1*, *10*). We used periodic quadratic arrays of such SRRs made from gold with the dimensions given by the electron micrographs shown in Fig. 1 and on the right-hand side of Fig. 2. All samples discussed have an area of 25 by 25 μm. For example, a lattice constant of a = 450 nm corresponds to a total number of 56 × 56 = 3136 SRRs (see supporting online material text for fabrication and characterization details).

Figure 2 summarizes a number of different spectroscopic results in the form of a "matrix." Electron micrographs of corresponding samples are shown together with illustrations of the two linear polarization configurations used in the measurements. All of the lattice constants (from 450 to 900 nm) are much smaller than the 3-μm LC-resonance wavelength, discussed in the experiments and calculations below.

After the illustration in Fig. 1, the magnetic field vector of the incident light has a vanishing component normal to the coil for normal incidence conditions. Thus, coupling to the LC resonance is only possible if the electric field vector has a component normal to the plates of the capac-

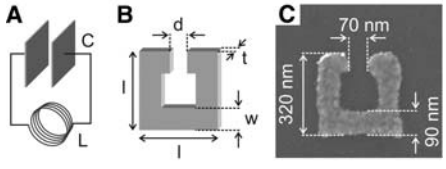


Fig. 1. Illustration of the analogy between a conventional LC circuit (**A**), consisting of an inductance L, a capacitance C, and the single SRRs used here (**B**). l, length; w, width; d, gap width; t, thickness. (**C**) An electron micrograph of a typical SRR fabricated by electron-beam lithography. The thickness of the gold film is t = 20 nm. For normal incidence, where the magnetic field vector $\boldsymbol{B}$ lies in the plane of the coil, the electric field vector $\boldsymbol{E}$ of the incident light must have a component parallel to the electric field of the capacitor to couple to the LC circuit. This allows the coupling to be controlled through the polarization of the incident light (Fig. 2).

[1]Institut für Nanotechnologie, Forschungszentrum Karlsruhe in der Helmholtz-Gemeinschaft, D-76021 Karlsruhe, Germany. [2]Institut für Angewandte Physik, Universität Karlsruhe (TH), Wolfgang-Gaede-Straße 1, D-76131 Karlsruhe, Germany. [3]Ames Laboratory and Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA. [4]Institute of Electronic Structure and Laser (IESL), Foundation for Research and Technology-Hellas (FORTH), 71110 Heraklion, Crete, Greece.

*To whom correspondence should be addressed. E-mail: soukoulis@ameslab.gov

itance (*8*), which corresponds to the left column of spectra in Fig. 2. For all lattice constants, *a*, two distinct resonances are clearly visible. Their spectral position does not depend on *a*. With increasing lattice constant *a*, the resonances narrow to some extent because of the reduced interaction between the SRRs, but their spectral position remains essentially unchanged, as expected for the electric and magnetic resonant responses of SRRs (*6*). The long-wavelength resonance around 3-μm wavelength completely disappears if the electric field vector is rotated by 90° (right column of spectra in Fig. 2). This is expected for the LC resonance, according to our above reasoning (Fig. 1). Also, see the more thorough theoretical discussion in (*8*). To further strengthen our interpretation of the 3-μm resonance in terms of an LC resonance, we show corresponding spectra for closed rings (*6*) rather than split rings in Fig. 2, G and H. Indeed, the 3-μm resonance does not occur for either linear polarization in this case, and the reflection and transmission spectra are nearly identical for the two polarizations, apart from minor deviations, because of imperfections in the nanofabrication process.

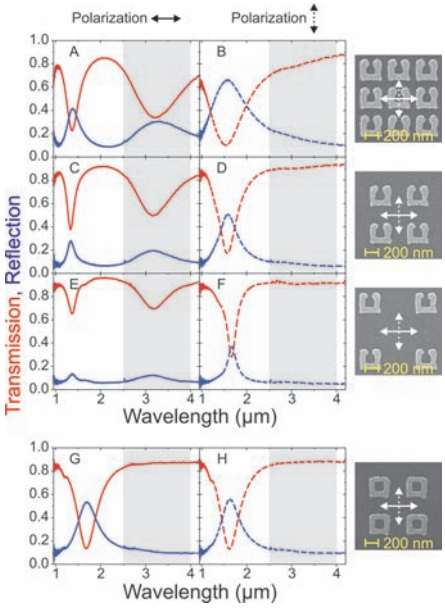


Fig. 2. Measured transmission (red) and reflection (blue) spectra. In each row of this "matrix," an electron micrograph of the sample is shown on the right-hand side. The two polarization configurations are shown on top of the two columns. In the first row (**A** and **B**), the lattice constant of the SRRs is a = 450 nm; in the second row (**C** and **D**), it is a = 600 nm; and in the third row (**E** and **F**), it is a = 900 nm. (A) to (F) correspond to nominally identical SRRs. In the last row (**G** and **H**), results for closed-ring resonators with a = 600 nm are shown. The combination of these spectra unambiguously shows that the resonance at about 3-μm wavelength (highlighted by the gray areas) is the LC resonance of the individual SSRs.

The additional transmission minimum between 1- and 2-μm wavelengths is due to the electric resonance arising from the currents induced by the electric field of the incident radiation in the metallic sides of the SRR. These currents are parallel to the polarization of the electric field. This electric resonance is related to the plasmon resonance in a thin continuous wire but shifted to nonzero frequency as a result of the additional depolarization field arising from the finite side length of the SRR. It appears independent of the excitation of the circular currents of the LC resonance and also occurs for the closed rings (Fig. 2, G and H). This electric resonance is sometimes also referred to as the particle-plasmon resonance. Finally, we have performed measurements (*11*) under an angle of up to 40° with respect to the surface normal, such that the magnetic field vector of the incident light acquires a component normal to the coils [similar to (*10*)]. As expected, the 3-μm resonance persists and does not shift. Larger angles of incidence are not possible in our experiments because of geometrical restrictions of the microscope used and because of the limited depth of focus.

To further strengthen our above assignment of the peaks and to determine the implications for possible left-handed materials, we compared these measurements with theory. In these calculations, the actual geometrical parameters (Fig. 1B) of the experiment were used. The calculations were performed with the software package CST Microwave Studio (Computer Simulation Technology GmbH, Darmstadt, Germany). The Drude model is used to describe the metal. That is, the effective permittivity of metals in the infrared spectral region is given by

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega + i\omega_c)}$$

where ω_p is the plasma frequency and ω_c is the collision frequency. For bulk gold, the parameters are $\omega_p = 2\pi \times 2.175 \times 10^{15}$ s^{-1} and $\omega_c = 2\pi \times 6.5 \times 10^{12}$ s^{-1} (*12*). For the thin gold films of our SRR, we expected that electrons experience additional scattering resulting from the metal surfaces. Thus, we used a value that is 1.65 times as large for the scattering frequency than in bulk. This obviously increases the absorption and improves the agreement between simulated and experimental observed losses. Numerical results are shown in Fig. 3, which can be directly compared with the experiment (Fig. 2). The overall agreement is very good. In particular, the absolute spectral positions of the peaks are reproduced with better than 10% accuracy. Furthermore, the polarization dependence (e.g., compare Fig. 3, A and B) also agrees with the experiment. Importantly, the resonance around 3-μm wavelength disappears for the closed rings (e.g., compare Fig. 3, C and G). According to our above reasoning, such behavior is indeed expected for the

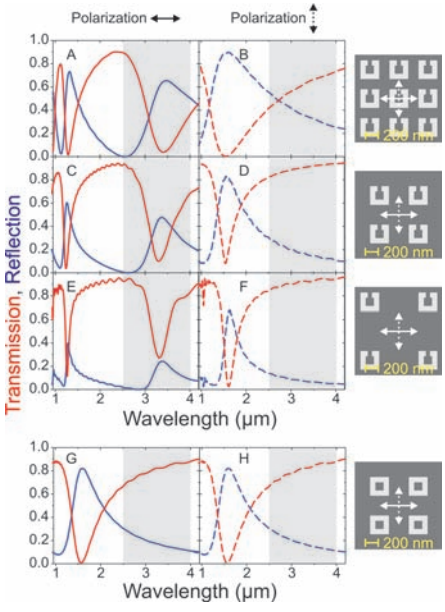


Fig. 3. (**A** to **H**) Calculated transmission (red) and reflection (blue) spectra, corresponding to the experiments shown in Fig. 2.

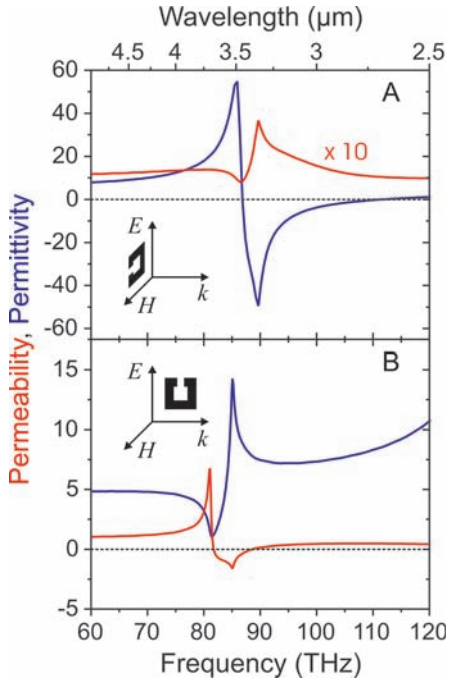


Fig. 4. The real part of the retrieved effective permeability μ and permittivity ε around the LC resonance of the SRR for the case of purely electric coupling (**A**) and purely magnetic coupling (**B**). In (A), μ has been multiplied by a factor of 10 to improve visibility. A negative μ region is observed for magnetic coupling (B). The resonance/antiresonance coupling between ε and μ is due to the periodic structure of the metamaterial (*14*). *E*, electric field vector; *H*, magnetic field vector; *k*, wave vector.

magnetic resonance, because the incident electric field cannot couple to the capacitance of the LC circuit. This argument is also true for the vertical polarization.

Figure 4 shows the effective permeability μ and the effective permittivity ε around the LC resonance, retrieved from the computed scattering data for two different orientations of the SRRs with respect to the incident wave (*13*). Figure 4A corresponds to the spectra shown on Fig. 3A, where only the electric field can couple to the LC resonance (*8*). In this case we obtain an electric resonant response in ε, accompanied by a simultaneous antiresonant behavior in μ (*14*). The retrieved data for the other polarization (Fig. 3B) exhibits no structure in ε and μ in this frequency range (*11*). In Fig. 4B, the beam configuration is such that the magnetic field can couple to the LC resonance, whereas the electric field cannot. For this polarization, a magnetic resonant response in μ is obtained with a negative value of μ in the 85-THz region. This is an important precondition for the realization of a metamaterial with a negative index of refraction. The retrieved ε exhibits an antiresonant behavior in this case (*14*).

Our results have two important consequences. First, usual ferromagnetic and antiferromagnetic resonances tend to die out above gigahertz frequencies. Thus, one can usually safely assume that the magnetic permeability of optical materials is unity. In other words, the optical properties of materials are exclusively determined by the optical polarization; the optical magnetization is zero. This is no longer true for the metamaterials presented here, enabling interesting new effects in linear optics as well as in nonlinear optics. Indeed, additional theoretical calculations show that the local fields within the gap of the LC circuit can be orders of magnitude larger than in free space or in bulk, which potentially enhances nonlinear effects and conversion efficiencies considerably. Second, a negative magnetic permeability would allow for negative-index materials at optical frequencies, which seemed totally out of reach just a few years ago.

References and Notes

1. R. A. Shelby, D. R. Smith, S. Schultz, *Science* **292**, 77 (2001).
2. V. G. Veselago, *Sov. Phys. Usp.* **10**, 509 (1968).
3. J. B. Pendry, *Phys. Rev. Lett.* **85**, 3966 (2000).
4. D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, S. Schultz, *Phys. Rev. Lett.* **84**, 4184 (2000).
5. D. R. Smith, S. Schultz, P. Markos, C. M. Soukoulis, *Phys. Rev. B* **65**, 195104 (2002).
6. T. Koschny, M. Kafesaki, E. N. Economou, C. M. Soukoulis, *Phys. Rev. Lett.* **93**, 107402 (2004).
7. C. Sirtori, *Nature* **417**, 132 (2002).
8. N. Katsarakis, T. Koschny, M. Kafesaki, E. N. Economou, C. M. Soukoulis, *Appl. Phys. Lett.* **84**, 2943 (2004).
9. D. R. Smith, J. B. Pendry, M. C. K. Wiltshire, *Science* **305**, 788 (2004).
10. T. J. Yen *et al.*, *Science* **303**, 1494 (2004).
11. S. Linden *et al.*, data not shown.
12. M. A. Ordal *et al.*, *Appl. Opt.* **22**, 1099 (1983).
13. D. R. Smith, S. Schultz, P. Markos, C. M. Soukoulis, *Phys. Rev. B* **65**, 195104 (2002).
14. T. Koschny, P. Markos, D. R. Smith, C. M. Soukoulis, *Phys. Rev. E* **68**, 065602 (2003).

15. We acknowledge the support by the Center for Functional Nanostructures (CFN) of the Deutsche Forschungsgemeinschaft (DFG) within project A.1.4. The research of M.W. is further supported by the DFG-Leibniz award 2000 and that of C.M.S. by the Alexander von Humboldt senior-scientist award 2002, by Ames Laboratory (contract no. W-7405-Eng-82), European Union Future and Emerging Technologies project, Development and Analysis of Left-Handed Metamaterials, and Defense Advanced Research Projects Agency (contract no. MDA 972-01-2-0016).

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1351/DC1
SOM Text

17 September 2004; accepted 20 October 2004

A Chiral Route to Negative Refraction

J. B. Pendry

Negative refraction is currently achieved by driving the magnetic permeability and electrical permittivity simultaneously negative, thus requiring two separate resonances in the refracting material. The introduction of a single chiral resonance leads to negative refraction of one polarization, resulting in improved and simplified designs of negatively refracting materials and opening previously unknown avenues of investigation in this fast-growing subject.

Negative refraction is an intriguing and counterintuitive phenomenon that has attracted much attention. Not only does light bend the "wrong" way at a normal/negative interface, but there are even more surprising properties, such as the ability to construct a "perfect" lens for which the resolution is limited not by the wavelength but by the quality of manufacture (*1*, *2*). Negative refraction never occurs in nature, and we rely on artificial materials, metamaterials, to realize the effect as discussed in (*3*). In this paper, I discuss the consequences of chirality and show that it offers an alternative to the present routes to negative refraction. I produce a practical design that is chiral, has many advantages, and exhibits novel properties.

In the original description of negative refraction (*4*), it was stated that when the electrical permittivity and magnetic permeability are both negative light bends the wrong way at an interface. It was only much later, with the ability to construct artificial metamaterials, that the properties could be realized (*5*–*7*): The original prescription for a subwavelength array of thin metallic wires combined with resonant metallic rings has been extensively investigated, and negative refraction at microwave frequencies has been confirmed by several investigators (*8*–*13*). Although referred to as "left-handed" materials, I stress that the sense in which this term was used has nothing to do with chirality. Therefore I prefer to use the expression "negatively refracting" to avoid confusion.

The nonchiral designs suffer some limitations. They use two sets of resonant structures, one for the electric and the other for the magnetic response, and these structures have to be very carefully designed to resonate in the same frequency range. Figure 1 shows a typical schematic band structure where it was assumed that

$$\mu < 0, \omega_1 > \omega > \omega_3$$
$$\varepsilon < 0, \omega_2 > \omega > \omega_4 \quad (1)$$

where μ is the magnetic permeability, ω is the frequency, and ε is the effective electric permittivity. The negatively dispersing band between ω_2 and ω_3 is responsible for a negative refractive index. The structure of the metamaterial is required to be as fine as possible so that the fields experience an effectively homogeneous material. The pres-

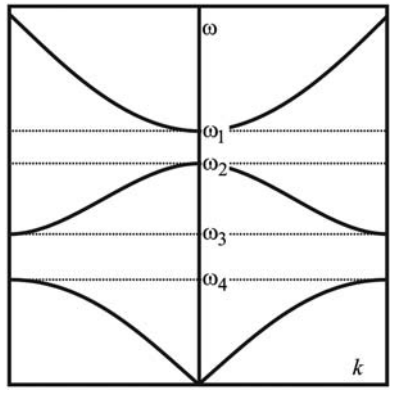


Fig. 1. Dispersion in a negatively refracting material: Typically two stop bands, $\omega_1 > \omega > \omega_2$ and $\omega_3 > \omega > \omega_4$, and a band of negative dispersion and hence of negative refraction, $\omega_2 > \omega > \omega_3$, are seen. In addition, there are two longitudinal modes (not shown), one magnetic in character and the other electric.

Department of Physics, Blackett Laboratory, Imperial College London, London SW7 2AZ, UK.

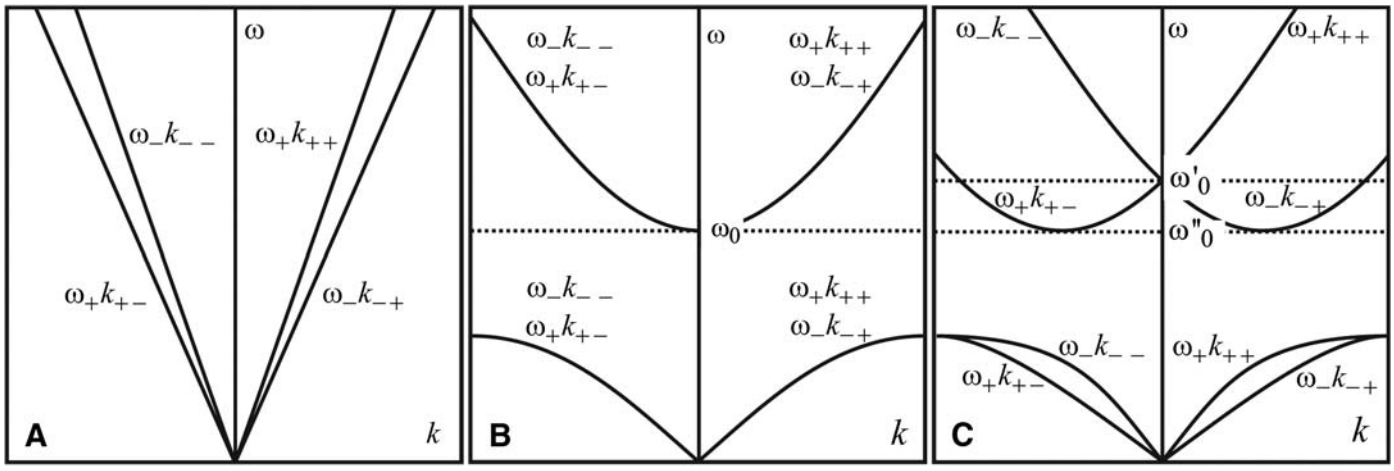


Fig. 2. (**A**) Dispersion of ω versus k in a homogeneous and isotropic chiral medium showing the two polarizations as nondegenerate. The subscripts on k refer first to the polarization and second to the sign of the group velocity. In this convention, polarization is positive if the projection of the photon spin on the z axis is positive. It does not refer to the projection of spin onto the wave vector. (**B**) Dispersion in a homogeneous and isotropic medium containing a set of resonant electric dipoles. Note the gap opened up where the permittivity is negative and the two degenerate transverse modes. There is an additional longitudinal mode, not shown, degenerate with the longitudinal modes at ω_0. (**C**) Introducing the resonant dipoles into a chiral medium splits the resonant transverse bands and results in a range of frequencies below ω_0' in which negative refraction can be seen for one of the polarizations. The longitudinal mode, again not shown, is degenerate with the transverse modes at ω_0'.

ent generation of designs go some way to achieving this but rarely do better than a wavelength-to-structure ratio of 10:1. Also, producing a magnetic resonance at optical frequencies will be particularly difficult if low loss is required.

Chiral materials exhibit a different refractive index for each polarization. The dispersion of wave vector, k, with ω is shown (Fig. 2A). Formally speaking we introduce a tensor, χ,

$$\chi_A = \begin{bmatrix} \chi_{EE} & \chi_{EH} \\ \chi_{HE} & \chi_{HH} \end{bmatrix} \quad (2)$$

which defines the response of the medium to an electromagnetic field:

$$\mathbf{D} = \chi_{EE}\mathbf{E} + \chi_{EH}\mathbf{H}$$
$$\mathbf{B} = \chi_{HE}\mathbf{E} + \chi_{HH}\mathbf{H} \quad (3)$$

where **D** is the electric displacement vector; **E**, the electric field intensity; **B**, the magnetic induction field; and **H**, the magnetic field intensity. The Supporting Online Material (SOM) Text relates χ to ω(k).

Next, consider another medium filled with resonant electric dipoles so that

$$\chi_B(\omega) = \begin{bmatrix} \left(1+\frac{\alpha^2}{\omega_0^2-\omega^2}\right) & 0 \\ 0 & 1 \end{bmatrix} \quad (4)$$

The response is shown in Fig. 2B. The resonance induces a band gap: a range of frequencies in which the electrical permittivity is negative and where there are no allowed states. The medium is not chiral, so the two transverse polarizations are degenerate. In addition to the transverse modes there is a single longitudinal mode, which is degenerate with the longitudinal modes at $k = 0$ and $\omega = \omega_0$.

Now consider what happens if dipole resonators are inserted into the chiral medium. The combined response is given by

$$\chi_C = \begin{bmatrix} \left(\chi_{EE} + \frac{\alpha^2}{\omega_0^2-\omega^2}\right) & \chi_{EH} \\ \chi_{HE} & \chi_{HH} \end{bmatrix} \quad (5)$$

and the schematic dispersion is plotted (Fig. 2C). Chirality splits the degenerate transverse modes and in doing so creates a range of frequencies just below ω_0' where the group velocity,

$$v_g = \partial\omega/\partial k \quad (6)$$

has the opposite sign to the phase velocity,

$$v_p = \omega/k \quad (7)$$

but only for one polarization. This is the signature of negative refraction. One can define a refractive index for each polarization,

$$n_\pm(\omega) = c_0 k_{\pm+}/\omega \quad (8)$$

where c_0 is the velocity of light in free space, and one of them is negative. An impedance can also be defined (SOM Text).

The bands shown in Fig. 2C are remarkable in several ways. In contrast to the nonchiral situation where there are two resonances and therefore two gaps, here there is only one gap. Hence, the transition to negative refraction is smooth and continuous at ω_0' with no gap. Only a chiral material can achieve this because it requires that the intersecting bands at $k = 0$ do not hybridize as they normally would do. Also, the bands at this point have finite group velocity but infinite phase velocity. For the chiral material, the minimum frequency, ω_0'', occurs as a finite wave vector, and hence the density of states diverges at this frequency.

In general when light in a vacuum is incident upon a slab of the new resonant chiral material, two refracted beams will be observed because a refracting surface can mix polarizations. However suppose the parameters are chosen such that

$$n_+ = -1 \quad (9)$$

and

$$Z_+ = +Z_0 \quad (10)$$

where Z_0 is the ratio of electric to magnetic fields for waves in the vacuum, that is, the impedance of the vacuum, and Z_+ is the ratio of electric to magnetic fields in the medium for the polarization that shows negative refraction (SOM Text). Then detailed calculations show that the slab is perfectly transparent to the + polarization. All the properties predicted for an isotropic nonchiral medium with

$$\varepsilon = -1, \ \mu = -1 \quad (11)$$

will be reproduced in this new medium but only for one of the polarizations. This includes the ability to focus the near fields and hence reproduce an image with resolution unlimited by wavelength.

This recipe for chiral negative refraction is a general one. Only the ingredients of a resonant system producing a band gap and chirality are needed. A moment's study of Fig. 2C will show that any minimum in ω(k) at $k = 0$ will produce negative refraction when split in this way.

Next I suggest a practical realization of a resonant chiral structure. I do this by

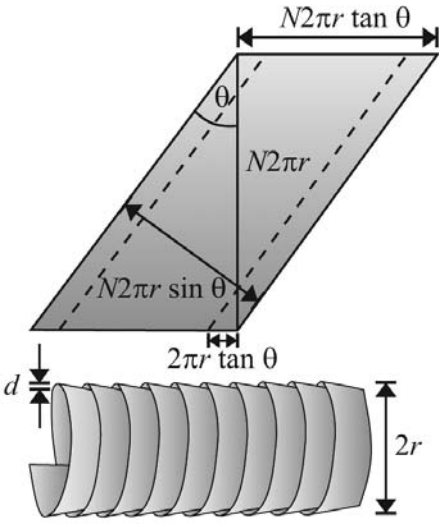


Fig. 3. Design for a chiral material. Continuous insulated strips of metal are wound in a helix and then individual coils are stacked in a three-dimensional log pile to make an isotropic structure.

winding a continuous insulated metal tape onto a cylinder so that it forms an overlapping helix (Fig. 3). More details of the performance of this structure are available (figs. S8 to S13). This is a chiral variant of the so-called Swiss roll structure used to produce negative permeability in the MHz range of frequencies (*14*). The structure is resonant because of inductance in the coiled helix and capacitance between the inner and outer layers of the helix. When current flows along the helix, not only does it produce a magnetic polarization along the axis, but it also produces an electric polarization because some of the current flows parallel to the axis. Typical values for the parameters are

$$r = 5 \times 10^{-3}\ \text{m}$$
$$a = 2 \times 10^{-2}\ \text{m},\ d = 1 \times 10^{-4}\ \text{m}$$
$$\theta = 5^\circ,\ N = 10 \qquad (12)$$

where a is the lattice constant of the log-pile structure and r, d, θ, and N are defined in Fig. 3. These parameters give negative refraction at around 100 MHz. The design can be tuned over a wide range of frequencies (SOM Text). Achieving strong chirality in the optical region of the spectrum is more difficult, but some promising design studies have been made (*15*).

The class of negatively refracting materials introduced here with the prescribed properties should open previously unknown avenues of investigation. Specific designs are greatly simplified with very compact internal structure, typically on a scale less than 1/100th of the free space wavelength at the resonant frequency. The structures offer further opportunities to extend the negative refraction concept.

References and Notes

1. J. B. Pendry, *Phys. Rev. Lett.* **85**, 3966 (2000).
2. J. B. Pendry, S. A. Ramakrishna, *J. Phys. Cond. Matter* **15**, 6345 (2003).
3. D. R. Smith, J. B. Pendry, M. C. K. Wiltshire, *Science* **305**, 788 (2004).
4. V. G. Veselago, *Soviet Physics Usp.* **10**, 509 (1968).
5. J. B. Pendry, A. J. Holden, W. J. Stewart, I. Youngs, *Phys. Rev. Lett.* **76**, 4773 (1996).
6. J. B. Pendry, A. J. Holden, D. J. Robbins, W. J. Stewart, *IEEE Trans. Microwave Theory Tech.* **47**, 2075 (1999).
7. D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, S. Schultz, *Phys. Rev. Lett.* **84**, 4184 (2000).
8. A. A. Houck, J. B. Brock, I. L. Chuang, *Phys. Rev. Lett.* **90**, 137401 (2003).
9. A. Grbic, G. V. Eleftheriades, *Phys. Rev. Lett.* **92**, 117403 (2004).
10. P. V. Parimi *et al.*, *Phys. Rev. Lett.* **92**, 127401 (2004).
11. E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopoulou, C. M. Soukoulis, *Nature* **423**, 604 (2003).
12. P. V. Parimi, W. T. Lu, P. Vodo, S. Sridhar, *Nature* **426**, 404 (2003).
13. E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopolou, C. M. Soukoulis, *Phys. Rev. Lett.* **91**, 207401 (2003).
14. M. C. K. Wiltshire *et al.*, *Science* **291**, 849 (2001).
15. Y. Svirko, N. Zheludev, *Appl. Phys. Lett.* **78**, 498 (2001).
16. I thank the Donostia International Physics Center (DIPC) for their hospitality during the course of this work; the Engineering and Physical Sciences Research Council for a senior fellowship; and the European Commission (EC) under project FP6-NMP4-CT-2003-505699, the U.S. Department of Defense/Office of Naval Research Multidisciplinary University Research Initiative grant N00014-01-1-0803, and the EC Information Societies Technology (IST) program Development and Analysis of Left-Handed Materials (DALHM) project number IST-2001-35511 for financial support.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5700/1353/DC1
SOM Text
Figs. S1 to S13

15 June 2004; accepted 1 September 2004

Hertz-Level Measurement of the Optical Clock Frequency in a Single $^{88}Sr^{+}$ Ion

H. S. Margolis,* G. P. Barwood, G. Huang, H. A. Klein, S. N. Lea, K. Szymaniec, P. Gill

The frequency of the 5s $^2S_{1/2}$–4d $^2D_{5/2}$ electric quadrupole clock transition in a single, trapped, laser-cooled $^{88}Sr^{+}$ ion has been measured by using an optical frequency comb referenced to a cesium fountain primary frequency standard. The frequency of the transition is measured as 444,779,044,095,484.6 (1.5) hertz, with a fractional uncertainty within a factor of 3 of that of the cesium standard. Improvements required to obtain a cesium-limited frequency measurement are described and are expected to lead to a $^{88}Sr^{+}$ optical clock with stability and reproducibility exceeding that of the primary cesium standard.

Accurate time and frequency measurement is a long-standing requirement of science and technology, with applications including the realization of the Système International (SI) base units of time and length, satellite-based navigation and ranging, precision measurements of fundamental constants, and tests of physical theories (*1*). Since 1967, the internationally agreed definition of the second has its basis in the ground-state hyperfine transition in the ^{133}Cs atom at 9,192,631,770 Hz, and Cs fountain primary frequency standards now have reproducibilities of around 1 part in 10^{15} (*1*). Recent studies have indicated the potential of optical frequency standards based on high-Q transitions in laser-cooled trapped ions or atoms to achieve even better stabilities and accuracies (*2*, *3*). Combined with optical frequency measurement techniques based on femtosecond combs (*4*, *5*), these transitions may be used as practical frequency standards generating a direct microwave output (*3*), raising the possibility of a future redefinition of the SI second. To assess the suitability of an optical standard as the basis for such a new definition, it is important to make accurate measurements of its frequency relative to the Cs standard and to evaluate its reproducibility. In the medium term, this would allow such standards to be used as secondary representations of the second, contributing to International Atomic Time (TAI).

An important class of optical frequency standards are those that have their basis in narrow-linewidth forbidden transitions in single laser-cooled trapped ions, such as $^{199}Hg^{+}$ (*2*), $^{171}Yb^{+}$ (*6*, *7*), $^{115}In^{+}$ (*8*), and $^{88}Sr^{+}$ (*9*, *10*). The best previously reported frequency measurements are for the $^{199}Hg^{+}$ standard at 282 nm (*11*) and the $^{171}Yb^{+}$ standard at 435 nm (*6*), with quoted uncertainties of 1 part in 10^{14}. We present a measure-

National Physical Laboratory (NPL), Teddington, Middlesex TW11 0LW, UK.

*To whom correspondence should be addressed. E-mail: helen.margolis@npl.co.uk

ment of the 5s $^2S_{1/2}$–4d $^2D_{5/2}$ optical clock transition frequency in $^{88}Sr^+$ at 674 nm, with a fractional uncertainty that is three times better than the previous best measurements.

A partial term diagram for the $^{88}Sr^+$ optical frequency standard is shown (Fig. 1A). The ion is trapped in an endcap trap (*12*) and laser-cooled to a few mK on the 5s $^2S_{1/2}$–5p $^2P_{1/2}$ transition at 422 nm. To prevent optical pumping into the metastable 4d $^2D_{3/2}$ state, we used 1092-nm laser radiation. Fluorescence from the 422-nm cooling transition is detected by a photomultiplier tube, and the micromotion of the ion is monitored and minimized in three dimensions by using radio frequency (rf)–photon correlation techniques (*13*, *14*). The reference for the optical frequency standard is the 5s $^2S_{1/2}$–4d $^2D_{5/2}$ electric quadrupole transition at 674 nm, which has a natural linewidth of 0.4 Hz. This is probed with the use of a 674-nm extended-cavity diode laser, which is stabilized to a high-finesse ultra-low-expansion (ULE) cavity by using the Pound-Drever-Hall technique (*15*), resulting in a laser linewidth of less than 100 Hz and a relative frequency stability of $1\times10^{-14}/\sqrt{\tau}$ for averaging times $\tau < 10$ s. The frequency-stabilized probe laser is shifted into resonance with the 674-nm clock transition frequency by using a double-passed acousto-optic modulator (AOM). The clock transition is observed by using the quantum jump technique (*16*); i.e., excitation to the metastable 4d $^2D_{5/2}$ level is detected from cessation of the strong 422-nm fluorescence from the cooling transition.

In a dc magnetic field (typically 1.4 μT), the clock transition splits into 10 Zeeman components (Fig. 1B). The $^{88}Sr^+$ ion is interrogated by using a computer-controlled sequence of operations during which the ion is alternately cooled at 422 nm and then probed at 674 nm. The probe laser pulses were typically 5 ms in duration, resulting in a Fourier transform–limited linewidth of about 200 Hz. The center frequency, ν_0, of the Zeeman structure is determined by using a four-point servo scheme with a typical cycle time of 15 to 20 s to probe a pair of Zeeman components, which are symmetrically placed around line center (*17*).

Absolute frequency measurements of the clock transition frequency were performed by simultaneously recording the AOM offset frequency and measuring the frequency of the light locked to the ULE cavity with the use of a femtosecond optical frequency comb. The setup of this frequency comb was similar to that described in (*9*), except that both the repetition rate and the carrier envelope offset frequency were stabilized to rf synthesizers referenced to the 10-MHz output of a hydrogen maser. The offset of the maser output frequency from 10 MHz was determined by comparison with the NPL Cs fountain (*18*) throughout the period of the frequency measurements. The stability of the rf synthesizers used on the comb contributes to the statistical uncertainty of the frequency measurements, but the comb linkage does not significantly affect the final systematic uncertainty.

The largest source of systematic frequency shift for the $^{88}Sr^+$ optical frequency standard arises from the electric quadrupole shift of the reference transition (*17*), which is due to the interaction between the electric quadrupole moment of the atomic states and any residual electric field gradient present at the position of the ion. After the treatment in (*19*), the resultant frequency shift of the 4d $^2D_{5/2}$ level with magnetic quantum number m_j' (j indicates total angular movement) is given by

$$\Delta\nu = \frac{3}{10h} Q_{dc}\Theta(D,5/2) \times \left(\frac{35}{12} - m_j'^2\right)(3\cos^2\beta - 1) \qquad (1)$$

where Q_{dc} is the residual dc quadrupole field gradient at the position of the ion, $\Theta(D,5/2)$ is the electric quadrupole moment of the 4d $^2D_{5/2}$ state, β is the angle between the principal axis of the quadrupole field gradient and the magnetic field axis, and h is Planck's constant. The 5s $^2S_{1/2}$ state has no electric quadrupole moment and so is not shifted by this effect. The residual dc quadrupole field gradient is determined from measurements of the trap secular frequencies and minimized by adjusting the dc voltages applied to the outer electrodes of the endcap trap (*17*). In this way the shift is reduced to the level of a few Hz. However, the quadrupole shift can be nulled by using one of two different techniques, with the result that the uncertainty in this shift can be reduced to a substantially lower level. In the first, referred to as method A, a particular pair of Zeeman components is selected, and frequency measurements are carried out for three mutually orthogonal orientations of the applied dc magnetic field. The average quadrupole shift for these three measurements is zero (*19*). In the second technique (method B), measurements are carried out by using three different pairs of Zeeman components that correspond to transitions with the three different possible values of $|m_j'|$, for example, pairs a, b, and c in Fig. 1B (*10*, *20*). From Eq. 1, the average quadrupole shift for these three transitions is again zero, independent of the magnetic field direction.

Frequency measurements of the 5s $^2S_{1/2}$–4d $^2D_{5/2}$ transition in $^{88}Sr^+$ were carried out on 11 separate days. The first 6 days of data (Fig. 2A) were taken with the use of method A to null the electric quadrupole shift, with measurements being carried out in three nominally orthogonal magnetic field directions corresponding to angles of β of about 11°, 101°, and 90°. The later 5 days of data (Fig. 2B) were taken with the use of method B to null the electric quadrupole shift. For 2 of these days, the magnetic field direction was such that $\beta \approx 11°$; for the other 3 days, β was about 90°. Each data point corresponds to typically 10,000 s of data, and the error bars represent statistical uncertainties only. The unweighted mean frequencies of the two sets of data (before correcting for systematic errors) are 444,779,044,095,485.0 (1.3) Hz and 444,779,044,095,486.2 (1.2) Hz, respectively, where the statistical standard uncertainties are shown in parentheses.

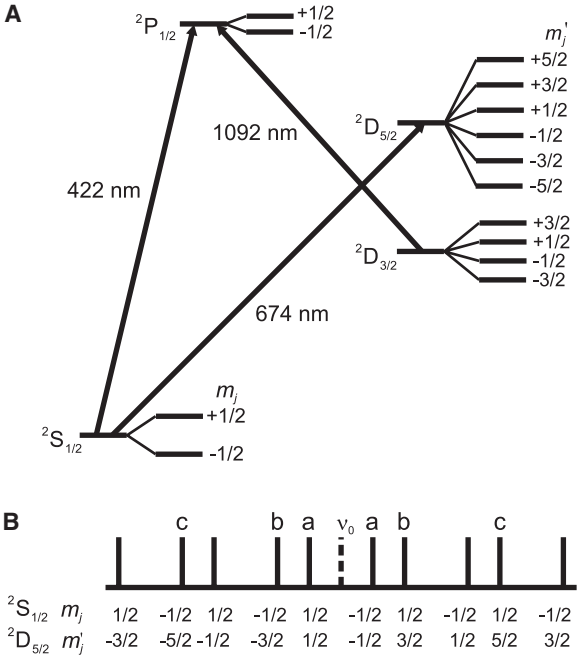


Fig. 1. (**A**) Partial term scheme for $^{88}Sr^+$ (nuclear spin $I = 0$), showing the transitions used to cool and probe the trapped ion. (**B**) Zeeman splitting of the 5s $^2S_{1/2}$–4d $^2D_{5/2}$ clock transition, showing the m_j quantum numbers for each component and the centroid ν_0 of the multiplet.

A detailed analysis of ion-related systematic errors for the $^{88}Sr^{+}$ optical frequency standard was presented in (*10*). The estimated sizes of these effects for our system are given in Table 1 together with some technical sources of systematic uncertainty, the effects of which have been evaluated experimentally wherever possible.

Although the quadrupole shift of the transition frequency is nominally nulled for both sets of data, the cancellation of the shift in method A requires the three magnetic field orientations to be exactly orthogonal. Allowing for an estimated 10° of uncertainty in the magnetic field directions chosen, the residual uncertainty due to the quadrupole shift is estimated to be around 10% of the mean magnitude of the shift. The cancellation of the quadrupole shift in method B relies only on the angle between the magnetic field axis and the quadrupole axis being stable over the course of a day's measurements, and so the residual uncertainty is assumed to be negligible in this case.

The thermal secular motion of the trapped ion and the micromotion at the trap drive frequency both lead to Stark shifts (which arise because the ion experiences a nonzero time-averaged value of the ac trapping field) and to second-order Doppler shifts. The shifts arising from the micromotion are estimated from the rf photon correlation signals (*14*), whereas those due to the secular motion are estimated from the intensity of the secular sidebands relative to the carrier. There is also a Stark shift arising from blackbody radiation, if the transition frequency is corrected to absolute zero. At 293 K, this is calculated to be 0.30 (8) Hz (*10*), with the uncertainty arising primarily from the uncertainty in the Stark shift coefficients.

The linear Zeeman effect is eliminated by probing two Zeeman components, which are symmetrically placed around line center. The quadratic Zeeman effect shifts both components in the same direction but, at the typical dc applied magnetic field of about 1.4 μT, is negligible compared to other sources of uncertainty. The blackbody radiation field, with a mean-square magnetic field amplitude of 7.01 μT^2 at 293 K, and ac magnetic fields similarly lead to negligible Zeeman shifts at the current level of precision.

The radiation used to cool and probe the trapped ion can cause ac Stark shifts of the clock transition frequency. The procedures used to null the electric quadrupole shift also null the tensor component of these ac Stark shifts, and so only the scalar components need be considered. The 674-nm ac Stark shift is negligible at the power level and beam waist size used in the trap. Although the 1092-nm and 422-nm beams are nominally switched off during the probe laser interrogation periods, in practice they are not perfectly extinguished. No significant frequency shift was detected between measurements carried out with and without an AOM switching the 1092-nm beam, setting a limit on the 1092-nm ac Stark shift under normal operating conditions from the measured extinction ratio of the AOM. Effective extinction of the 422-nm radiation is more important, because the cooling and probe transitions share the ground state as a common level. This beam is therefore blocked by using a combination of an AOM and a mechanical shutter, giving a total extinction ratio of 10^6. The residual 422-nm ac Stark shift due to light scattered around the shutter blade was determined by increasing the rf power leakage to the AOM by 15 dB, when it was nominally switched off, and scaling the frequency shift measured under these conditions to the normal leakage level.

Drifts in the ULE cavity frequency can lead to servo errors in the lock to the center of the Zeeman structure. To quantify this effect, we induced deliberate changes in both the magnitude and the direction of the cavity drift rate by adjusting the temperature control system, leading to larger-than-normal imbalances between the quantum jump rates on each side of the line. The relationship between measured frequency and quantum jump imbalance determined in this way was consistent with the relationship predicted on the basis of the observed linewidth and servo parameters. For the special case where measurements of the quantum jump rate are made at two frequencies separated by the linewidth of the Zeeman component, this relationship is given in (*21*).

The uncertainty in the maser reference frequency has contributions both from the

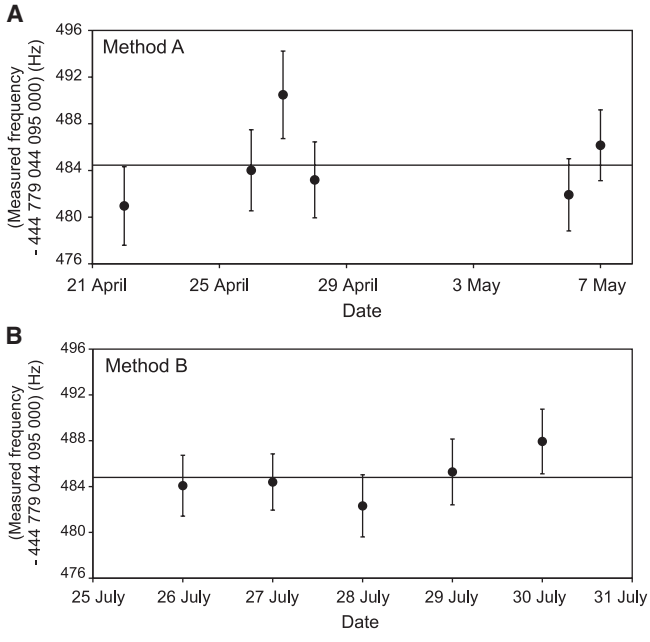


Fig. 2. Frequency of the $5s\ ^2S_{1/2}$–$4d\ ^2D_{5/2}$ transition in $^{88}Sr^{+}$, determined from (**A**) the average of measurements made in three nominally orthogonal magnetic field directions and (**B**) the average of measurements carried out by using three different pairs of Zeeman components corresponding to transitions for which $|m_j'| =$ 1/2, 3/2, and 5/2. The solid lines show the average measured value using each technique. Corrections have been applied for systematic frequency shifts as itemized in Table 1.

Table 1. Estimated size and standard uncertainty (Hz) of all systematic frequency shifts larger than 10 mHz.

Source	Method A		Method B	
	Shift	Uncertainty	Shift	Uncertainty
Quadrupole shift	0	0.5	0	<0.01
2nd-order Doppler shift due to micromotion	<0.01	0.01	<0.01	0.01
2nd-order Doppler shift due to secular motion	<0.01	0.01	<0.01	0.01
Stark shift due to micromotion	+0.01	0.01	+0.01	0.01
Stark shift due to secular motion	<0.01	0.01	<0.01	0.01
Blackbody Stark shift	+0.30	0.08	+0.30	0.08
1092-nm ac Stark shift	0	0.02	0	0.02
422-nm ac Stark shift	+1.4	0.8	+1.4	0.8
Servo errors	–1.0	0.6	–0.4	0.3
Maser reference frequency	0	0.7	0	0.7
Gravitational shift	0	0.1	0	0.1
Total estimated systematic shift	+0.7	1.3	+1.3	1.1

uncertainty in the Cs fountain frequency (1 part in 10^{15} statistical over the time scale of these measurements and 1 part in 10^{15} systematic) (*18*) and from the uncertainty in correcting for the frequency shifts in the cables used to transfer the maser signal to the femtosecond comb laboratory. Lastly, the uncertainty in the relative altitudes of the Sr ion trap and the Cs fountain leads to a small uncertainty because of the gravitational shift.

Correcting for the systematic shifts gives frequency values of 444,779,044,095,484.3 (1.9) Hz and 444,779,044,095,484.8 (1.6) Hz for the data taken with use of the two different methods of nulling the electric quadrupole shift, which agree to well within their statistical uncertainties. The unweighted mean of the two values gives a final value for the 674-nm electric quadrupole clock transition frequency in $^{88}Sr^{+}$ of 444,779,044,095,484.6 (1.5) Hz. In calculating the final uncertainty, only the statistical error is reduced, because the systematic uncertainties are mostly common to the two measurements. This result is in good agreement with earlier, less-accurate measurements (*9*, *10*) of the Sr clock transition frequency. It is also a factor of 3 more accurate than any previously reported optical frequency measurement, and its fractional uncertainty of 3.4×10^{-15} is within a factor of 3 of that of the NPL primary Cs standard (when both statistical and systematic errors are considered). Apart from the uncertainty arising from the Cs standard, the dominant sources of uncertainty in our measurement are technical in nature and can be reduced by refinements to the experimental arrangement. In particular, improvements to the extinction of the cooling laser radiation during the probe laser periods (e.g., by placing an iris after the shutter) will reduce the 422-nm ac Stark shift, whereas reductions in the probe laser linewidth and ULE cavity drift rate will reduce servo errors. A second endcap trap is also being developed, which will enable a more detailed investigation of systematic errors by means of two-trap comparisons. With these improvements, we anticipate a frequency measurement that is limited by the accuracy of the Cs fountain. As well as being of interest for a possible future redefinition of the second, measurements of this and other optical frequency standards over timescales of a few years will provide increasingly sensitive laboratory tests of the time invariance of fundamental constants (*22*, *23*).

References and Notes

1. P. Gill, Ed., *Proceedings of the 6th Symposium on Frequency Standards and Metrology* (World Scientific, Singapore, 2002).
2. R. J. Rafac *et al.*, *Phys. Rev. Lett.* **85**, 2462 (2000).
3. S. A. Diddams *et al.*, *Science* **293**, 825 (2001); published online 12 July 2001 (10.1126/science.1061171).
4. T. Udem, J. Reichert, R. Holzwarth, T. W. Hänsch, *Opt. Lett.* **24**, 881 (1999).
5. S. A. Diddams *et al.*, *Phys. Rev. Lett.* **84**, 5102 (2000).
6. J. Stenger, C. Tamm, N. Haverkamp, S. Weyers, H. R. Telle, *Opt. Lett.* **26**, 1589 (2001).
7. P. J. Blythe *et al.*, *Phys. Rev. A* **67**, 020501(R) (2003).
8. T. Becker *et al.*, *Phys. Rev. A* **63**, 051802(R) (2001).
9. H. S. Margolis *et al.*, *Phys. Rev. A* **67**, 032501 (2003).
10. A. A. Madej, J. E. Bernard, P. Dubé, L. Marmet, R. S. Windeler, *Phys. Rev. A* **70**, 012507 (2004).
11. T. Udem *et al.*, *Phys. Rev. Lett.* **86**, 4996 (2001).
12. C. A. Schrama, E. Peik, W. W. Smith, H. Walther, *Opt. Commun.* **101**, 32 (1993).
13. I. Siemers, M. Schubert, R. Blatt, W. Neuhauser, P. E. Toschek, *Europhys. Lett.* **18**, 139 (1992).
14. D. J. Berkeland, J. D. Miller, J. C. Bergquist, W. M. Itano, D. J. Wineland, *J. Appl. Phys.* **83**, 5025 (1998).
15. R. W. P. Drever *et al.*, *Appl. Phys. B* **31**, 97 (1983).
16. H. Dehmelt, *IEEE Trans. Instrum. Meas.* **31**, 83 (1982).
17. G. P. Barwood, H. S. Margolis, G. Huang, P. Gill, H. A. Klein, *Phys. Rev. Lett.* **93**, 133001 (2004).
18. K. Szymaniec, W. Chalupczak, P. B. Whibberley, S. N. Lea, D. Henderson, in preparation.
19. W. M. Itano, *J. Res. Natl. Inst. Stand. Technol.* **105**, 829 (2000).
20. P. Dubé, L. Marmet, A. A. Madej, J. E. Bernard, presentation at the 2004 Conference on Precision Electromagnetic Measurements, London, 27 June to 2 July 2004.
21. J. E. Bernard, L. Marmet, A. A. Madej, *Opt. Commun.* **150**, 170 (1998).
22. M. Fischer *et al.*, *Phys. Rev. Lett.* **92**, 230802 (2004).
23. E. Peik *et al.*, *Phys. Rev. Lett.* **93**, 170801 (2004).

24. We thank W. Chalupczak, D. Henderson, P. Stacey, and P. Whibberley for their contributions to the caesium fountain and hydrogen maser operation. This work was supported by the Department of Trade and Industry National Measurement System Length program under contract LE02/A01.

21 September 2004; accepted 25 October 2004

Multifunctional Carbon Nanotube Yarns by Downsizing an Ancient Technology

Mei Zhang,[1] Ken R. Atkinson,[2] Ray H. Baughman[1]*

By introducing twist during spinning of multiwalled carbon nanotubes from nanotube forests to make multi-ply, torque-stabilized yarns, we achieve yarn strengths greater than 460 megapascals. These yarns deform hysteretically over large strain ranges, reversibly providing up to 48% energy damping, and are nearly as tough as fibers used for bulletproof vests. Unlike ordinary fibers and yarns, these nanotube yarns are not degraded in strength by overhand knotting. They also retain their strength and flexibility after heating in air at 450°C for an hour or when immersed in liquid nitrogen. High creep resistance and high electrical conductivity are observed and are retained after polymer infiltration, which substantially increases yarn strength.

Archaeological evidence from the late Stone Age indicates that humans long ago discovered the basic secrets of spinning (*1*). Similar processes involving the twisting of centimeter-long fibers to make continuous yarns are critically important for many of today's industries, and they remain a focus of research and development. The present work reduces the diameters of fibers used for spinning by a factor of 1000, increases twist by about the same factor, and discovers useful properties for the resulting spun carbon nanotube yarns.

Nanotube spinning is motivated in part by interest in the very high strength and electrical and thermal conductivities of individual nanotubes (*2*). Breakthroughs have been made in wet spinning of single-walled nanotubes (SWNTs) (*3*–*6*) and in dry-state spinning of multiwalled nanotubes (MWNTs) and SWNTs (*7*, *8*), but the highest strength achieved with any of these spinning methods is about an order of magnitude lower than the strength of individual SWNTs, ~37 GPa (*2*).

There is no single best solution to the challenge of converting available nanotube powders into useful fibers and yarns. Excellent fiber strength (4.2 GPa) and modulus (167 GPa) have been achieved by incorporating SWNTs in a high-strength, high-modulus polymer, but electrical and thermal conductivities are low because of limitations on nanotube content (*5*). Much higher conductivities result for thermally annealed, solution-spun yarns comprising only SWNTs (*2*, *4*), but achieved mechanical properties are far lower than can be obtained using a polymer matrix for intertube stress transfer (*6*). Coagulation-spun yarns comprising ~60 weight % SWNTs in a polymer matrix have high strength (~1.8 GPa) and more than 10 times the toughness (~600 J/g) of any synthetic polymer, but the electrical conductivity is so low that charge/discharge rates for fiber-based supercapacitors are limited (*6*). The challenge is to produce yarns that are at the same time strong, creep resistant, highly

[1]NanoTech Institute and Department of Chemistry, University of Texas at Dallas, Richardson, TX 75083, USA. [2]CSIRO Textile & Fibre Technology, P.O. Box 21, Belmont, Victoria 3216, Australia.

*To whom correspondence should be addressed. E-mail: ray.baughman@utdallas.edu

conducting, and reversibly deformable over relatively large strains to absorb energy.

It is likely that the characteristics of the frictional forces that couple fibers in twist-based yarns change as fiber diameter decreases to the nanoscale. Nevertheless, a generic equation (*9*) provides useful insights for spinning nanotube yarns. Specifically, the ratio of yarn tensile strength (σ_y) to the tensile strength of the component fibers (σ_f) is approximately

$$\sigma_y/\sigma_f \approx \cos^2 \alpha[1 - (k \operatorname{cosec} \alpha)] \qquad (1)$$

where $k = (dQ/\mu)^{1/2}/3L$, α is the helix angle that fibers make with the yarn axis, d is the fiber diameter, μ is the friction coefficient between fibers, L is the fiber length, and Q is the fiber migration length (i.e., the distance along the yarn over which a fiber shifts from the yarn surface to the deep interior and back again).

The $\cos^2 \alpha$ term in Eq. 1 describes the strength decrease of a twisted assembly of continuous fibers, which occurs because the fibers in the twisted yarn are inclined at the angle α with respect to the tensile axis. For short fibers, however, in the absence of twist there is little strength because there are no significant transverse forces to bind the fiber assembly together. The $[1 - (k \operatorname{cosec} \alpha)]$ term describes the generation of transverse forces by transfer of the tensile load to the yarn surface, which locks the fibers together as a coherent structure. The components of k show that the strength obtainable for a given level of twist increases with increasing coefficient of friction and fiber length and with decreasing fiber diameter and fiber migration length.

Like the spinning method of Jiang *et al*. (*7*), the present process involves drawing carbon nanotubes from a MWNT "forest" (in which all nanotubes grow from a substrate in a manner resembling tree trunks in a dense bamboo forest and have nearly the same height). Li *et al*. (*8*) made the important discovery that MWNT yarns with much higher strengths could be directly spun from aerogels during nanotube synthesis by chemical vapor deposition (CVD). Although a stable twisted yarn was shown (*8*), the ratio of nanotube length (~ 30 μm) to yarn diameter was very low (about unity), which should limit the benefits of twisting.

Scanning electron microscope (SEM) images (Fig. 1, A and B) show yarn assembly during our spinning process, in which MWNTs ~10 nm in diameter are simultaneously drawn from the MWNT forest and twisted. The direction of drawing was orthogonal to the original nanotube direction and parallel to the substrate, although the spinning process is sufficiently robust that the angle between the initial nanotube direction and the draw direction can be decreased from 90° to almost 0°. The nanotube forests were grown on an iron catalyst–coated substrate by CVD, using a method (*10*) that builds on important advances of the Dai (*11*) and Ren (*12*) groups. Although this spinning process is amenable to automation for spinning continuous yarns, the present results are for yarns that were hand-drawn from a nanotube forest while they were twisted with a variable-speed motor operating at ~2000 rpm (*13*). In some spinning trials, the spun yarns had such small diameters (~ 1 μm) that spinning was somewhat like hauling in a fish with an invisible line—with the additional complication that the line is simultaneously being twisted at a high rate. The achieved yarn length was limited to ~1 m by the arm length of the person doing the drawing, although there is no fundamental limit on achievable yarn lengths.

Although the MWNT lengths are 0.1 to 1% those of staple fibers such as wool and cotton, this simple method results in yarns that are strong, very flexible, tough, and highly conducting. The geometry of the spun yarn resembles that of wool or cotton yarn, and we were able to select spinning parameters by extending the classic theory of staple yarns from the microscale to the nanoscale (*9*), using additional insights from theoretical work on nanotubes by the Pipes and Ruoff groups (*14*, *15*).

The yarn diameter was set by controlling the width of the forest sidewall that was used to generate an initial wedge-shaped ribbon, which is shown converging from about the thickness of the forest to that of the yarn at the apex (Fig. 1) (fig. S2). The forest sidewall width ranged from less than 150 μm to ~3 mm; the resulting unplied yarn diameters were between ~1 and 10 μm. A forest sidewall 200 μm wide produced a twisted yarn ~2 μm in diameter, and a forest area of 1 cm^2 could generate an estimated 50 m of this yarn. The inserted twist was typically ~80,000 turns/m, versus ~1000 turns/m for a highly twisted conventional textile yarn with diameter 80 times as large. Because $\tan \alpha/\pi D$ is the twist required to provide the helix angle α for a yarn having diameter D, such huge twists provide helix angles comparable to those of highly twisted conventional textiles.

Measurements of yarn diameter by SEM and yarn mass per length showed that the twisted yarn had a density of ~0.8 g cm^{-3}. The linear density of the unplied yarns (called "singles") was typically ~10 μg m^{-1}, compared with the usual 10 mg m^{-1} and 20 to 100 mg m^{-1} for cotton and wool yarns, respectively. About 100,000 MWNTs pass through the cross section of a nanotube yarn 5 μm in diameter, as compared with the 40 to 100 fibers in the cross section of typical commercial wool (worsted) and cotton yarns.

SEM images of singles, two-ply, and four-ply MWNT yarns, respectively, are shown in Fig. 2, A to C. The two-ply yarns were obtained by overtwisting a singles yarn and then allowing it to relax (untwist) around itself until it reached a torque-balanced state. This procedure was repeated for the two-ply yarn, using an opposite twist direction, to produce four-ply yarns; such a method was used to make rope in the days of sailing vessels. The two-ply structure was torque balanced, as indicated by MWNT alignment along the yarn axis (Fig. 2B).

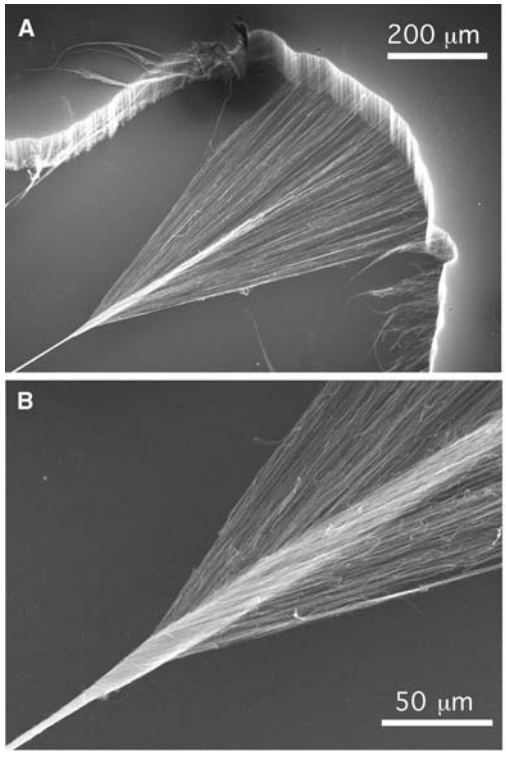


Fig. 1. (**A** and **B**) SEM images, at two different magnifications, of a carbon nanotube yarn in the process of being simultaneously drawn and twisted during spinning from a nanotube forest outside the SEM. The draw twist process was interrupted and the sample was transferred to a SEM for image recording. A forest strip (width ~600 μm) formed the pictured twisted yarn, 3.2 μm in diameter (fig. S2). The MWNTs, ~10 nm in diameter with length 10,000 times their width, form small bundles of a few nanotubes each in the forest, with individual nanotubes moving in and out of different bundles. The three-dimensional connectivity caused by intermittently switched bundling, visible in SEM micrographs, is believed to be important for the spinning process.

Unlike singles yarn of a conventional textile, the MWNT singles yarn largely retains twist when the yarn ends are released. The enhanced locking of twist possibly reflects the high interfiber contact area per yarn volume as a result of the very high surface-to-volume ratio of the MWNTs. Particularly surprising is the observation that twist is retained up to the location of the break point for singles yarns that have been broken by tensile extension. In part because of twist retention, highly twisted singles yarns can be knitted and knotted (Fig. 2, D and E). Abrasion and knotting, especially with an overhand knot, seriously degrades the strength of most polymer fibers (including those used for antiballistic vests), conventional textile yarns, individual polymer chains, actin filaments, and DNA, causing rupture at the entrance to the knot (*16*, *17*). This is not the case for the investigated singles and two-ply nanotube yarns, where tensile failure is not observed in the vicinity of an inserted overhand knot. High abrasion resistance is suggested by the absence of ultimate tensile failure in a long yarn loop that was pulled through a very tight overhand knot. This useful resistance to knot-induced failure is likely a consequence of the very small tensile and compressive strains introduced by bending a MWNT 10 nm in diameter to the radius of a yarn knot; these strains are 0.05% of those that would occur in a fiber 20 μm in diameter with the same bending radius.

Although very high electrical conductivities have been observed for yarns of pure nanotubes (*4*, *7*, *8*), continuous composite fibers involving nanotubes and insulating polymers generally have low conductivities, whether the polymer is present during the spinning process or added after spinning. We find that the intertube mechanical coupling introduced by twisting largely maintains electronic connectivity between nanotubes during infiltration (*18*) of poly(vinyl alcohol) (PVA), thereby avoiding a major decrease in electrical conductivity. The investigated yarns (diameters from 2 to 10 μm) had a four-probe electrical conductivity of ~300 S/cm at room temperature and a negative temperature dependence of resistance (~ –0.1% per °C between liquid nitrogen and room temperature). PVA infiltration decreased yarn electrical conductivity by only ~30%, leading to MWNT/PVA composite yarns whose electrical conductivity is more than 150 times that of coagulation-spun nanotube composite fibers (*6*) containing this insulating polymer.

The untwisted yarns were so weak that they broke when pulled away from surfaces that they accidentally contacted, whereas the singles yarns had measured tensile strengths between 150 and 300 MPa (*13*). Higher strengths, between 250 and 460 MPa, were observed for two-ply yarns. These values are engineering stresses based on cross-sectional area measured by SEM for the unstressed yarn. As a result of a giant Poisson's ratio effect, the true stress at break (normalized to the true cross-sectional area at the beginning of yarn rupture) is ~30% larger. From the maximum observed density of 0.8 g cm^{-3}, we calculate the density-normalized failure stress of the two-ply yarns to be between 310 and 575 MPa/g·cm^{-3}, as compared with values between 50 and 500 MPa/g·cm^{-3} for polymer-free nanotube fiber spun from an aerogel (*8*) and >3 GPa/g·cm^{-3} for the highest performance graphite fibers (*19*). Infiltration with PVA (*18*) increased the observed strengths of singles yarns to 850 MPa.

The pure nanotube yarns had a much larger strain-to-failure (up to 13%) than graphite fibers (~1%). This high failure strain, combined with high failure strength, meant that the work needed to break the yarns (called toughness) was also high: ~14 J/g for the singles yarn, 20 J/g for the two-ply yarn, and 11 J/g for the PVA-infiltrated singles yarns, which combined their higher strength with a lower strain-to-failure (~3 to 4%). Although this toughness for the two-ply yarn (20 J/g) is above that of graphite fibers (12 J/g) and approaches that of commercial fibers used for antiballistic vests (~33 J/g for Kevlar fibers), far greater toughness has been demonstrated (*6*) for solution-spun SWNT/PVA composite fibers (600 J/g).

The stress-strain curves (Fig. 3A) show rapid collapse of the yarn stress once the breaking strain is exceeded, as also seen for conventional yarns (*9*). A simple model of tensile failure is thought to apply: Initial failure of the MWNTs in a central zone increases the freedom of movement of the MWNTs in the surrounding zones, which reduces transverse forces and ultimately leads to sliding of nanofibers with respect to each other (drafting) and catastrophic failure of the yarn. This process ensures that the ends of the broken yarn are drafted, as we observed.

The nanotube yarns showed hysteretic stress-strain curves when subjected to load-unload cycles (Fig. 3B). Although complete unloading did not return the yarn to its original length, the initial hysteresis loop was essentially unshifted on subsequent cycles. Depending on initial strain, the observed energy loss per stress-strain cycle of a two-ply MWNT yarn was 9 to 22% for cycle strain of 0.5%, 24 to 28% for cycle strain of 1.5% (Fig. 3C), and 39 to 48% for the maximum reversible cycle strain (2 to 3% for total strains up to 8%). Within a hysteresis loop, the effective modulus on initial unloading and initial reloading was much larger than for the final parts of the unloading and reloading steps (Fig. 3D). Also relevant for applications, the failure strength of nanotube yarn (singles and two-ply) was unaffected by 50 loading-unloading cycles over a stress range of 50% of the failure stress. The nanotube yarns were resistant to creep and associated stress relaxation: The stress relaxed no more than 15% when a two-ply nanotube yarn was held for 20 hours at 6% strain (initial

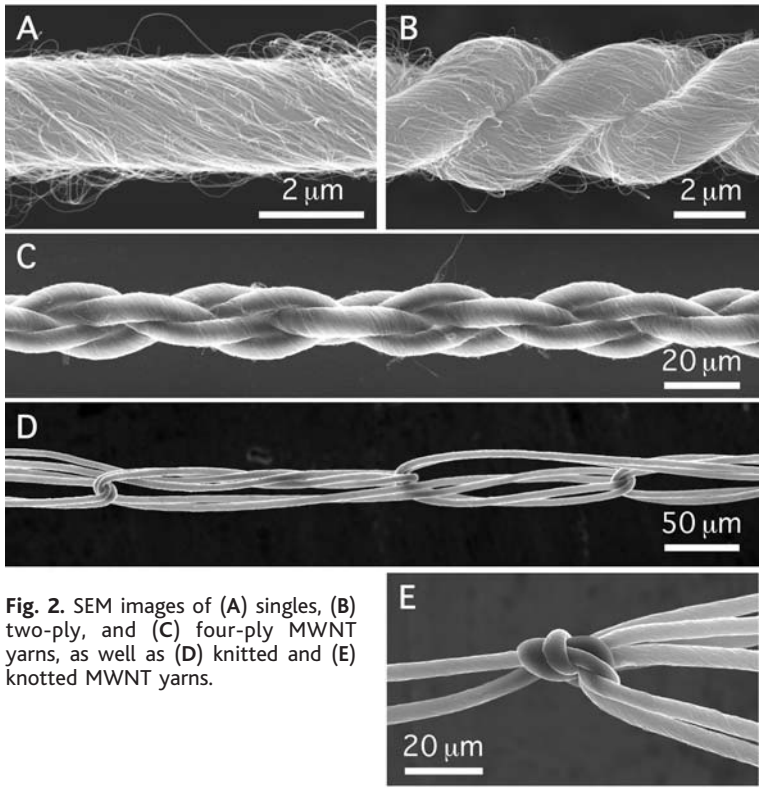


Fig. 2. SEM images of (**A**) singles, (**B**) two-ply, and (**C**) four-ply MWNT yarns, as well as (**D**) knitted and (**E**) knotted MWNT yarns.

stress 170 MPa), and this small stress relaxation occurred within the first 20 min and was largely viscoelastic. Thermal stability was also high: The failure strength of a two-ply yarn (300 MPa) was essentially unchanged after heating in air at 450°C for an hour. Although air oxidation was evident in SEM micrographs, a nanotube yarn held at 450°C for 10 hours was sufficiently strong and flexible to be tightly knotted. Tight knot tying was also possible while the nanotube yarn was immersed in liquid nitrogen.

We observed giant Poisson's ratios for the nanotube yarns, which increased with increasing strain from 2.0 to 2.7 for MWNT singles yarn and from 3.3 to 4.2 for two-ply yarn (Fig. 3E). This Poisson's ratio of 4.2 means that elongating the yarn by a strain ε provides a strain of −4.2 ε in each of the lateral dimensions and a fractional volume decrease of 7.4 ε, versus a fractional volume increase of ~0.4 ε for an ordinary solid with a typical Poisson's ratio of ~0.3.

The origin of these giant Poisson ratios is topological and is basically the same as for "finger trap" toys, selenium and tellurium single crystals, and the helically wound muscular hydrostats that provide stiffness for squid tentacles and the bodies of various worms (*20*). Unwinding an ideal helix by stretching provides a Poisson's ratio of >0.5 and stretch densification as long as the helix angle is sufficiently small (<54.73°). Hearle (*21*) long ago predicted giant Poisson's ratios for conventional yarns when this helix angle α is small and the ratio of fiber tensile modulus to yarn bulk modulus is large, and Poisson's ratios of up to 0.8 have been predicted by Pipes and Hubert (*14*) for twisted nanotube arrays. These Poisson's ratios, and the associated stretch-induced densification of up to 7.4 ε, might be used for tuning the absorption and permeability of nanotube yarns and textiles by applying small applied strains in a yarn direction.

The MWNT yarns are interesting as multifunctional materials. Their strength, toughness, mechanical energy damping capability, and resistance to knot-induced failure could be exploited, as could yarn diameters that are 2% the diameter of a human hair. Replacing metal wires in electronic textiles with these nanotube yarns could provide important new functionalities, such as the ability to actuate as an artificial muscle and to store energy as a fiber supercapacitor or battery. The small yarn diameters, like those of microdenier yarns used for soft fabrics, could eliminate the uncomfortable rigidity sometimes found for metal wire–containing conducting textiles that provide radio or microwave absorption, electrostatic discharge protection, textile heating, or wiring for electronic devices.

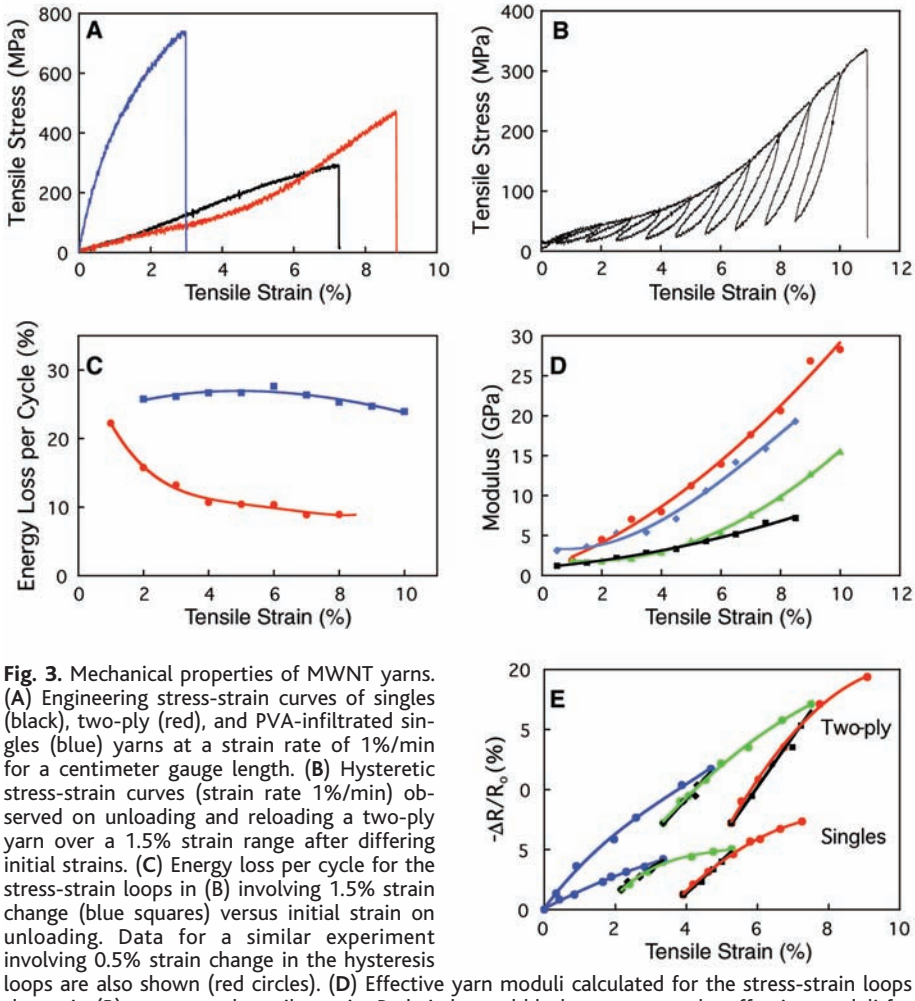


Fig. 3. Mechanical properties of MWNT yarns. (**A**) Engineering stress-strain curves of singles (black), two-ply (red), and PVA-infiltrated singles (blue) yarns at a strain rate of 1%/min for a centimeter gauge length. (**B**) Hysteretic stress-strain curves (strain rate 1%/min) observed on unloading and reloading a two-ply yarn over a 1.5% strain range after differing initial strains. (**C**) Energy loss per cycle for the stress-strain loops in (B) involving 1.5% strain change (blue squares) versus initial strain on unloading. Data for a similar experiment involving 0.5% strain change in the hysteresis loops are also shown (red circles). (**D**) Effective yarn moduli calculated for the stress-strain loops shown in (B) versus total tensile strain. Red circles and black squares are the effective moduli for the beginning and end of unloading, respectively; blue diamonds and green triangles are those for the beginning and end of reloading, respectively. (**E**) Percent change in diameter and length of a two-ply yarn (top) and a singles yarn (bottom) measured during yarn stretching in a SEM. Symbols: blue circles, initial stretch; black diamonds, first stress decrease; green circles, second stress increase; black squares, second stress decrease; red circles, stress increase until yarn rupture. Curves in (C) to (E) are guides for the eye.

References and Notes

1. E. J. W. Barber, *Prehistoric Textiles* (Princeton Univ. Press, Princeton, NJ, 1992).
2. R. H. Baughman, A. A. Zakhidov, W. A. de Heer, *Science* **297**, 787 (2002).
3. B. Vigolo *et al.*, *Science* **290**, 1331 (2000).
4. L. M. Ericson *et al.*, *Science* **305**, 1447 (2004).
5. S. Kumar *et al.*, *Macromolecules* **35**, 9039 (2002).
6. A. B. Dalton *et al.*, *Nature* **423**, 703 (2003).
7. K. Jiang, Q. Li, S. Fan, *Nature* **419**, 801 (2002).
8. Y. Li, I. A. Kinloch, A. H. Windle, *Science* **304**, 276 (2004).
9. J. W. S. Hearle, P. Grosberg, S. Backer, *Structural Mechanics of Fibers, Yarns, and Fabrics*, vol. 1 (Wiley, New York, 1969).
10. Aligned nanotube forests comprising MWNTs ~8 to ~15 nm in diameter were synthesized in a quartz tube 45 mm in diameter by atmospheric-pressure CVD of 5 mol % C_2H_2 in He at 680°C, at a total flow rate of 580 sccm for 10 min. The catalyst was an iron film, 5 nm thick, that was deposited on a Si wafer substrate by electron beam evaporation. SEM and thermal gravimetric measurements indicated that the purity of the spun yarns was very high (~96 to 98% C in the form of MWNTs), with 2 to 4% Fe and amorphous carbon. No carbon particles were observed.
11. S. Fan *et al.*, *Science* **283**, 512 (1999).
12. Z. F. Ren *et al.*, *Science* **282**, 1105 (1998).
13. See supporting data on *Science* Online.
14. R. B. Pipes, P. Hubert, *Compos. Sci. Technol.* **62**, 419 (2002).
15. D. Qian, W. K. Liu, R. S. Ruoff, *Compos. Sci. Technol.* **63**, 1561 (2003).
16. A. M. Saitta, P. D. Soper, E. Wasserman, M. L. Klein, *Nature* **399**, 46 (1999).
17. Y. Arai *et al.*, *Nature* **399**, 446 (1999).
18. MWNT/PVA composite yarns were made either by soaking a singles yarn for 15 hours in 5 wt % aqueous PVA solution or by passing a singles yarn through a drop of this solution during spinning, and then drying. The molecular weight of the PVA was in the range 77,000 to 79,000, and it was 99.0 to 99.8% hydrolyzed.
19. J.-B. Donnet, *Carbon Fibers* (Dekker, New York, 1998).
20. R. H. Baughman, S. Stafström, C. Cui, S. O. Dantas, *Science* **279**, 1522 (1998).
21. J. W. S. Hearle, *J. Polym. Sci. C* **20**, 215 (1967).

22. We thank the staff of the University of Texas at Dallas NanoTech Institute and clean room for their assistance. Supported by Defense Advanced Research Projects Agency/U.S. Army Research Office grant W911NF-04-1-0174, Texas Advanced Technology Program grant 009741-0130-2003, and the Robert A. Welch Foundation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1358/DC1
Materials and Methods
Figs. S1 and S2

19 August 2004; accepted 22 October 2004

Water-Assisted Highly Efficient Synthesis of Impurity-Free Single-Walled Carbon Nanotubes

Kenji Hata,*† Don N. Futaba,* Kohei Mizuno, Tatsunori Namai, Motoo Yumura, Sumio Iijima

We demonstrate the efficient chemical vapor deposition synthesis of single-walled carbon nanotubes where the activity and lifetime of the catalysts are enhanced by water. Water-stimulated enhanced catalytic activity results in massive growth of superdense and vertically aligned nanotube forests with heights up to 2.5 millimeters that can be easily separated from the catalysts, providing nanotube material with carbon purity above 99.98%. Moreover, patterned, highly organized intrinsic nanotube structures were successfully fabricated. The water-assisted synthesis method addresses many critical problems that currently plague carbon nanotube synthesis.

Single-walled carbon nanotubes (SWNTs) are a key aspect in the emerging field of nanotechnology; however, large-scale synthesis is still limited because of the difficulties in synthesizing SWNTs. Current synthesis methods suffer from the production of impurities that must be removed through purifications steps, which can damage the nanotubes. Dispersion of SWNTs in solutions for further processing also presents challenges because the smooth-sided tubes readily aggregate and form parallel bundles or ropes as a result of van der Waals interactions. We report a rational yet simple and general synthetic approach that concurrently addresses these problems, in which the activity and lifetime of the catalysts are dramatically enhanced by the addition of a controlled amount of water vapor in the growth atmosphere.

We wanted to find a weak oxidizer that would selectively remove amorphous carbon but would not damage the nanotubes at the growth temperature, because coating of the catalyst particles by amorphous carbon during chemical vapor deposition (CVD) reduces their activity and lifetime (*1*). We found that water acts in promoting and preserving catalytic activity. SWNTs were grown by ethylene CVD by using Ar or He with H_2 that contained a small and controlled amount of water vapor (*2*). Balancing the relative levels of ethylene and water was crucial to maximize catalytic lifetime. Water-assisted growth was successfully carried out on various catalysts that generate SWNTs, including Fe nanoparticles (*3*) from $FeCl_3$ and sputtered metal thin films (Fe, Al/Fe, Al_2O_3/Fe, Al_2O_3/Co) on Si wafers, quartz, and metal foils, which demonstrates the generality of our approach.

Water-stimulated catalytic activity results in the growth of dense and vertically aligned SWNT forests with millimeter-scale height in a 10-min growth time. Our best result to date is 2.5 mm in 10 min (Fig. 1, A and B). In contrast with standard ethylene CVD growth, where the catalysts are only active for about 1 min, a height increase of the forests has been observed after 30 min for water-assisted growth. The SWNT/catalyst weight ratio exceeds 50,000%, more than 100 times as high as that of the high-pressure carbon monoxide (HiPco) process (*4*). Provided that the amount of water is well controlled, growths are highly reproducible.

A close examination (Fig. 1C) at the ledge of the SWNT forest illustrates that the nanotubes are densely packed and vertically aligned from the substrate. Low-resolution transmission electron microscopy (TEM) studies (Fig. 1D) of the as-grown forest reveal the presence of only thin nanotubes and the absence of metallic particles and supporting materials that usually comprise a major constituent of as-grown material. High-resolution TEM studies (Fig. 1E and fig. S1) show that the nanotubes are clean SWNTs free from amorphous carbon and metal particles. We have taken hundreds of high-resolution TEM images, and double- or multi-walled carbon nanotubes (MWNTs) were rarely found. Raman spectra (fig. S2) at 514 nm excitation showed clear radial breathing mode peaks (RBM), which confirmed the existence of SWNTs. The sizes of the SWNTs were estimated from the peaks to be in the range of 1 to 3 nm, in agreement with those measured by TEM.

The SWNT forest structure can be easily removed from the substrate with, for example, a razor blade (movie S1). After removal, the substrate is still catalytically active and can grow SWNT forests again, indicating a root-growth mode and the presence of the catalysts on the substrate. Thermo-gravimetric analysis (TGA) was implemented on 10 mg of the as-grown material (Fig. 2A). No measurable residue remained after heating above 750°C, indicating very high purity. The combustion range of the SWNTs was 550°C to 750°C, with the peak weight reduction occurring at 700°C, a result very similar to that of purified, high-quality SWNTs synthesized by a laser-oven method (*5*). Quantitative elemental analysis with x-ray fluorescence spec-

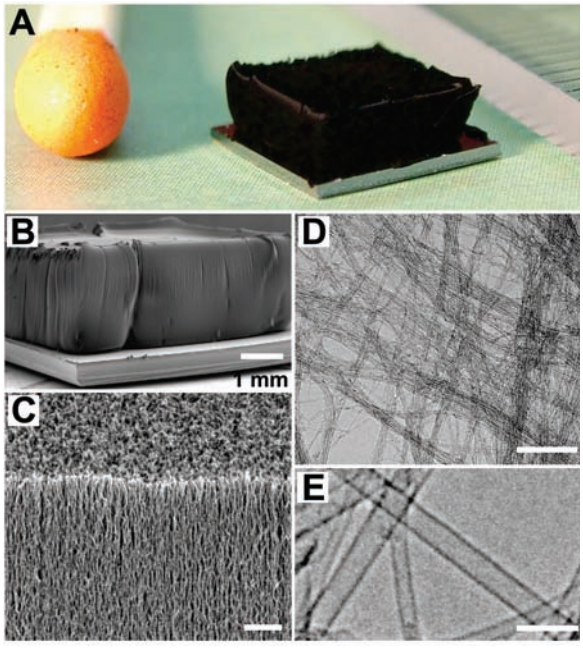


Fig. 1. SWNT forest grown with water-assisted CVD. (**A**) Picture of a 2.5-mm-tall SWNT forest on a 7-mm by 7-mm silicon wafer. A matchstick on the left and ruler with millimeter markings on the right is for size reference. (**B**) Scanning electron microscopy (SEM) image of the same SWNT forest. Scale bar, 1 mm. (**C**) SEM image of the SWNT forest ledge. Scale bar, 1 μm. (**D**) Low-resolution TEM image of the nanotubes. Scale bar, 100 nm. (**E**) High-resolution TEM image of the SWNTs. Scale bar, 5 nm.

Research Center for Advanced Carbon Materials, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, 305-8565, Japan.

*These authors contributed equally to this work
†To whom correspondence should be addressed. E-mail: kenji-hata@aist.go.jp

trometry detected 0.013% Fe as the only impurity, meaning carbon purity was more than 99.98%.

Pure SWNTs should allow for the investigation of the intrinsic magnetic properties of SWNTs and for the study of the photoluminescent properties that usually requires dissolution of the SWNTs. For example, a two-dimensional excitation-emission contour plot spectrum (Fig. 3) from preliminary spectrofluorimetric measurements on the as-grown SWNT forest clearly shows discrete spectral peaks. Each peak corresponds to characteristic absorption-emission wavelengths from van Hove optical transitions of SWNTs and can be assigned to specific SWNT structures (*6*). These characteristic peaks have been observed only in individual SWNTs in aqueous micellar suspensions in the past (*6*), and thus the observation of the fluorimetric peaks in the as-grown SWNT forest strongly indicates that the SWNTs are not heavily bundled but that many individual SWNTs exist. Peak locations in the spectrofluorimetric contour plot were mapped out and compared with that of the HiPco tubes (*6*), where a coincidence in peak location was found, which provides additional evidence that the tubes are SWNTs. However, the individual peak intensities differ noticeably from those of the HiPco tubes [e.g., the absence of the peak at 1105-nm emission and 647-nm adsorption wavelengths with assigned tube index (7.6)], which indicates a different abundance distribution of nanotube radii and chiralities. Moreover, a detailed investigation reveals that the spectrofluorimetric contour plot of the SWNT forest is richer in structure than that of the HiPco tubes; we tentatively attribute this result to a wider distribution of nanotubes in our samples. Our results provide a direct route to map the detailed composition of the as-grown SWNT materials and can be adapted to directly study the dependence of the nanotube distribution on the synthesis conditions and catalysts without any ambiguity.

Several additional points and experiments regarding the effect of water deserve comment. First, TGA (Fig. 2B) on pure SWNT material using N_2 gas with water shows that SWNT oxidization starts at about 950°C, which indicates that water does not oxidize and damage SWNTs at the growth temperature. We believe that the small initial weight increase is due to physisorption, supported by the fact that the weight returns to its initial value by subsequent annealing in dry N_2 gas. Second, a black amorphous carbon-coated quartz tube was cleaned transparent by flowing water vapor at 750°C, which provides direct evidence of water-induced oxidation of amorphous carbon. In the literature, hot water on carbon nanotubes has been used to purify amorphous carbon (*7*, *8*). Also, it has been reported that amorphous carbon is effectively removed from MWNTs by introducing water vapor into the CVD furnace (*9*). Metal particles further stimulate oxidation of carbon [for example, carbon in contact with metal particles is removed at temperatures as low as 225°C (*5*)], and thus this effect should further assist the role of water as a protective agent against amorphous carbon coating. These results suggest that water-assisted growth could be applied to other growth systems, such as methane and acetylene CVD, or to grow other nanotubes, such as MWNTs. We believe that our understanding concerning the role of water represents a reasonable basic description, although future work is required to quantify and better understand the effect of water.

Realization of large-scale organized SWNT structures of desired shape and form is important for obtaining scaled-up functional devices. With the assistance of water, SWNTs grow easily from lithographically patterned catalyst islands into well-defined vertical-standing organized structures, as demonstrated by the large-scale arrays of macroscopic cylindrical pillars (Fig. 4A) with 150-μm radius, 250-μm pitch and a

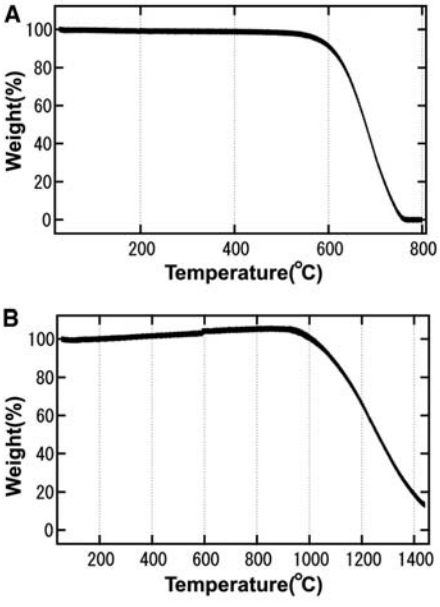


Fig. 2. Thermogravimetric properties of the SWNT material. (**A**) TGA data (ramp rate, 10°C/min) of a 10-mg sample of the SWNT material in air. (**B**) TGA data (ramp rate, 10°C/min) of a 9-mg sample of the SWNT material in N_2 (flow rate, 100 cc/min. standard temperature and pressure) passed through a water bubbler.

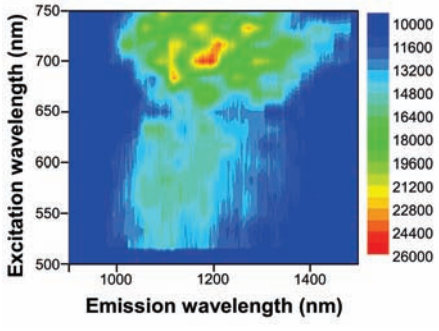


Fig. 3. Contour plot of fluorescence intensity versus excitation and emission wavelengths for the as-grown SWNT forest sample.

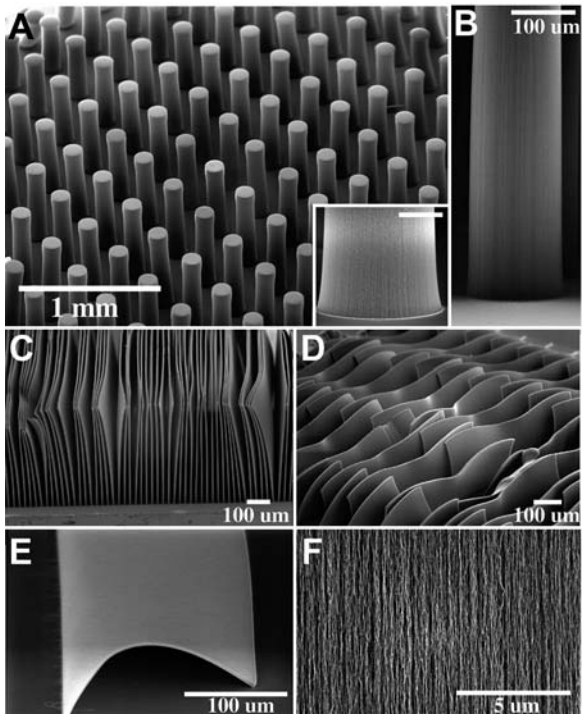


Fig. 4. SEM images of organized SWNT structures. (**A**) SEM image of SWNT cylindrical pillars with 150-μm radius, 250-μm pitch, and ~1-mm height. Inset, SEM image of a root of a pillar. Scale bar, 50 μm. (**B**) Side view of a pillar. Scale bar, 100 μm. (**C** and **D**) SEM images of SWNT sheets 10 μm thick. (**E**) SEM image of an isolated SWNT sheet 5 μm thick. (**F**) SEM image of the sheet face.

height close to 1 mm. A close examination (Fig. 4B) shows that the pillars are standing vertically from the substrate. Notably, the cross section of the SWNT structure corresponds well with the patterned catalyst (inset of Fig. 4A), and thus it is possible to fabricate arbitrary shapes of organized SWNT structures in which the base is lithographically defined and the height is controlled by the growth time. To further explore this unusual opportunity, we templated rows of catalytic stripe patterns and succeeded in growing pseudo two-dimensional organized SWNT structures (Fig. 4, C and D) that resemble sheets. A close investigation of a sheet face (Fig. 4F) reveals that the SWNTs are well aligned, with high uniformity. Some of these sheets are curved like pages in a book, which demonstrates their flexibility. This aspect is highlighted in Fig. 4E, in which an isolated thin SWNT sheet 5 μm thick was fabricated. Although this sheet formed a well-organized structure, its flexibility allowed it to bow and touch the surface, a point that suggests these thin sheets could be arbitrarily laid down, for example, by mechanical forces, gas flows, or electric fields.

Our approach is applicable to other synthesis methods developed for the mass production of SWNTs, such as rotary kiln, floating catalyst, and fluidized bed, addressing simultaneously such critical problems as scalability, purity, and cost. Thus, our approach represents an advance toward a realization of large-scale SWNT material. Additionally, our SWNTs are pure enough for use in various fields ranging from biology and chemistry to magnetic research. Highly pure SWNTs could be grown into scaled-up macroscopic organized structures with defined shape, be it a three-dimensional complex structure or a two-dimensional flexible sheet; potential applications include optical polarizers and field-emitter arrays for flat-panel displays.

References and Notes

1. S. Helveg *et al.*, *Nature* **427**, 426 (2004).
2. Material and methods are available as supporting material on *Science* Online.
3. H. Choi *et al.*, *Nano Lett.* **3**, 157 (2003).
4. M. Cinke *et al.*, *Chem. Phys. Lett.* **365**, 69 (2002).
5. W. I. Chiang *et al.*, *J. Phys. Chem. B* **105**, 8297 (2001).
6. M. J. O'Connell *et al.*, *Science* **297**, 593 (2002).
7. K. Tohji *et al.*, *Nature* **383**, 679 (1996).
8. K. Tohji *et al.*, *J. Phys. Chem. B* **101**, 1974 (1997).
9. A. Cao, X. Zhang, C. Xu, D. Wu, B. Wei, *J. Mater. Res.* **16**, 3107 (2001).
10. We thank Y. Kakudate for x-ray analyses, T. Yokoi for assistance with spectrofluorimetric measurements, K. Suenaga, K. Urita for some TEM observations, and T. Okazaki and M. Yudasaka for helpful discussions. Partial support by the New Energy and Industrial Technology Development Organization (NEDO) Nano Carbon Technology project and the use of the AIST Nano-Processing Facility are acknowledged.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1362/DC1
Materials and Methods
Figs. S1 and S2
Movie S1

7 September 2004; accepted 21 October 2004

Mars' South Polar Ar Enhancement: A Tracer for South Polar Seasonal Meridional Mixing

A. L. Sprague,[1]* W. V. Boynton,[1] K. E. Kerry,[1] D. M. Janes,[1] D. M. Hunten,[1] K. J. Kim,[2] R. C. Reedy,[2] A. E. Metzger[3]

The gamma ray spectrometer on the Mars Odyssey spacecraft measured an enhancement of atmospheric argon over southern high latitudes during autumn followed by dissipation during winter and spring. Argon does not freeze at temperatures normal for southern winter (~145 kelvin) and is left in the atmosphere, enriched relative to carbon dioxide (CO_2), as the southern seasonal cap of CO_2 frost accumulates. Calculations of seasonal transport of argon into and out of southern high latitudes point to meridional (north-south) mixing throughout southern winter and spring.

Between autumn (areocentric longitude of the Sun L_s 0° to 90°) (*1*) and winter in the southern hemisphere of Mars, about 25% of the atmosphere accumulates as a thick southern polar cap of CO_2 frost. Argon, a noncondensable gas, is left behind in the polar region (along with N_2, O_2, and CO) and becomes more enriched relative to CO_2, the main constituent in the atmosphere, as autumn progresses. The atmosphere near the poles tends to be isolated from the equatorial regions because of the conservation of angular momentum. If winds attempt to flow toward a pole, they are turned in the direction of the planet's rotation and form a vortex. This phenomenon has received attention in connection with the terrestrial polar ozone holes, the chemistry of which is connected to the isolation of the winter polar stratosphere and the special chemistry that takes place in this cold dark region. On Mars, very low temperatures above the southern polar winter night were discovered from analysis of Infrared Temperature Mapper measurements made from the Viking Orbiters (*2*, *3*). A substantial depletion in CO_2 might be the cause of the localized very low temperatures. Such a depletion would be accompanied by a large enhancement of the noncondensables Ar and N_2, and these were suggested to be enhanced by as much as a factor of 20 (*4*), particularly if CO_2 depletion was the cause of the localized cold spots. Here, we describe measurements of Ar in the polar region of the southern hemisphere.

Two of the three instruments comprising the gamma ray spectrometer (GRS) on Mars Odyssey (*5*–*7*) can study CO_2 frost accumulation and the cumulative effects of the sum of all noncondensables at southern latitudes by measuring the count rates of thermal neutrons, which show slightly lower values as noncondensables accumulate in the atmosphere (*8*). The GRS has the ability to measure Ar alone from the flux of the 1294-keV gamma ray associated with the 110-min half-life decay of Ar following neutron capture (*9*) [see supporting online material (SOM) text for details of this emission]. The instrument has a circular footprint on Mars with a diameter equal to about 240 km (*10*). We compute the fractional content of Ar relative to the total atmosphere in the polar region [mass mixing ratio (f_{GRS})] by dividing the mass of Ar measured by the GRS over the polar area by the mass of CO_2 in the polar atmosphere as predicted by the National Aeronautics and Space Administration Ames Research Center Mars Global Circulation Model (MGCM) run 2002.17 (*11*, *12*).

The relative Ar abundance (f_{GRS}) over Mars' southern polar latitudes from 75° to 90°S peaked at L_s 98°, 193 solar days (*13*) after CO_2 frost accumulation had begun (Fig. 1). The Ar abundance then decreased continuously throughout winter even though CO_2 frost accumulation continued. A minimum in Ar mass mixing ratio occurred after solid CO_2 began to sublime off the cap in early spring. Although the increase and decrease in Ar abundance are notable, the real surprise is that the data indicate transport of Ar equatorward throughout winter and poleward in spring. The Ar mixing ratio drops below the Viking Lander 2 (VL2) value (*14*) but is still measurable, despite rapid dilution

[1]Lunar and Planetary Laboratory, 1629 East University Boulevard, University of Arizona, Tucson, AZ 85721–0092, USA. [2]Institute of Meteoritics, MSC03-2050, University of New Mexico, Albuquerque, NM 87131–0001, USA. [3]Jet Propulsion Laboratory, Pasadena, CA 91109, USA.

*To whom correspondence should be addressed. E-mail: sprague@lpl.arizona.edu

by CO_2 subliming off the southern seasonal cap. This indicates mixing of atmosphere from mid-latitudes where the Ar mass mixing ratio is at the normal seasonal level.

The GRS data permit calculation of horizontal mixing coefficients that MGCMs cannot calculate with certainty. Horizontal mixing coefficients are derived with the use of Ar as a tracer, much as vertical eddy mixing coefficients (*15*) for the terrestrial stratosphere are derived with the use of methane as a minor constituent tracer. The derived mixing coefficients give insights into the rate of transport that would account for the observations and provide measured quantitative constraints for GCMs that seek to explain the past, present, and future meteorological conditions at Mars. The meridional mixing we infer from our data is useful for fine tuning the windfields used in the model calculations and for providing important constraints on GCMs attempting to predict martian meteorology.

The mass of Ar measured by the GRS was compared with a model prediction that assumes all Ar left behind by condensation of CO_2 during autumn and winter remains confined in the polar region (Fig. 2). The model shows what the Ar abundance would be if condensation flow (advection) were the only mechanism at work to transport Ar into and out of the polar area. Ar is transported into the polar region by the bulk atmosphere flowing toward the winter polar region. In the model, CO_2 freezes onto the seasonal cap and Ar is left behind and accumulates until the end of winter. There is a steep decrease at the onset of spring from subliming CO_2 ice diluting and carrying Ar away in the wind directed toward the equator. This advection-only model is computed with the CO_2 mass in the atmosphere and on the ground as predicted by the MGCM (see SOM text for advection-only model details). The amount of Ar ultimately left over the polar region in the model depends on the depth of the CO_2 frost in the southern cap, according to the MGCM prediction. The depth of CO_2 frost precipitation has also been measured by the three instruments of the GRS (*7*, *16*, *17*) and found to be within 10% agreement by season.

The Ar measured by the GRS is a few percent less than that predicted by the advection-only model during autumn. At the onset of winter, the difference increases with measured Ar that is lower than the model prediction by a factor of 8 to 9. The decrease continues until Ar reaches a minimum in early spring. The uncertainty in the measurements is small (typically less than 3%). In mid-spring the Ar abundance increases again (Fig. 2). This behavior indicates that another mechanism must be at work to remove Ar from the polar atmosphere during winter and to bring Ar back to the polar region during spring.

We have formulated a transport equation to explore the measured net flux of Ar into and out of the southern polar region and to compare the amount of Ar carried by condensation flow with that carried by fluctuations in the mean wind. The measured net flux of Ar, F_{Ar}, can be expressed as

$$F_{Ar} = v_x \rho f_{Ar} - K_x \rho \left(\frac{df_{Ar}}{dx} \right) \qquad (1)$$

where ρ is the mean seasonally adjusted ambient atmospheric mass density, predicted by the MGCM; f_{Ar} is the Ar mass mixing ratio appropriate for the location and season in the mid-latitude ambient atmosphere as computed in the MGCM; df_{Ar}/dx is the gradient in the mass mixing ratio over the distance through which Ar is mixed into and out of the GRS footprint; v_x is the north-south wind speed (*18*); and K_x is the eddy mixing parameter. K_x is solved for by using the other known or computed quantities. Our computations define the positive x direction to be toward the southern pole.

F_{Ar}, on the left-hand side of the equation, is computed from the measurements of the GRS. The first term on the right-hand side of Eq. 1 is the flux of Ar carried by condensation flow. This Ar is entrained in the atmosphere (primarily CO_2) moving toward the pole from low latitudes as the southern seasonal cap forms in autumn and winter. This condensation flow constitutes a small wind from the north with velocity v_x sustained throughout the period of CO_2 frost formation from L_s 0° to 180°. In spring, the first term becomes the flux of entrained Ar moving toward the equator as the southern seasonal CO_2 cap rapidly sublimes. The wind direction is from the south and v_x becomes negative in the computations.

The condensation flow is offset by a second process, at work in autumn and winter, to remove Ar from the high-latitude region, even as it is accumulating. This is expressed in the second term on the right-hand side of Eq. 1, which describes the flux of Ar into and out of the southern polar regions controlled by the gradient in the mixing ratio and by deviations from the mean winds. This term is usually called eddy transport or eddy mixing. Another process of

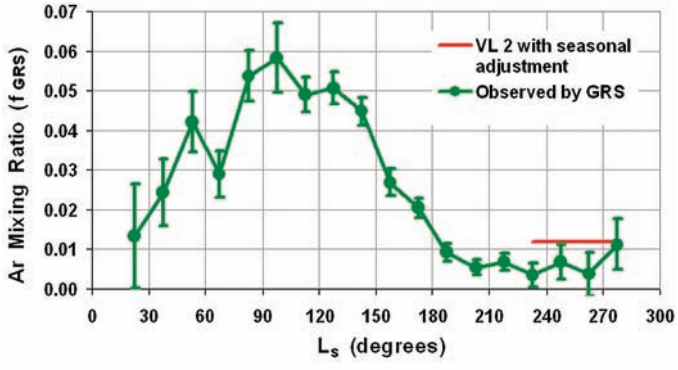


Fig. 1. Argon mass mixing ratios over Mars' southern polar latitudes (75° to 90°S) during autumn, winter, and spring, as determined from measurements by the GRS on the Mars Odyssey spacecraft at 15° increments in L_s (*1*). Error bars indicate 1σ limits on counting statistics. An additional 5 to 10% uncertainty in the internal calibration of the GRS is not shown but would result in the ensemble of points uniformly at either slightly higher or lower mixing ratio. The low point at L_s 65° is marginally outside the level of uncertainty and could represent temporal fluctuation in the amount of Ar occurring during that period. The red horizontal line corresponds to the Ar mass mixing ratio (0.0145) measured by the Viking Lander 2 (VL 2) gas chromatograph mass spectrometer, which when adjusted for season is decreased to 0.0119.

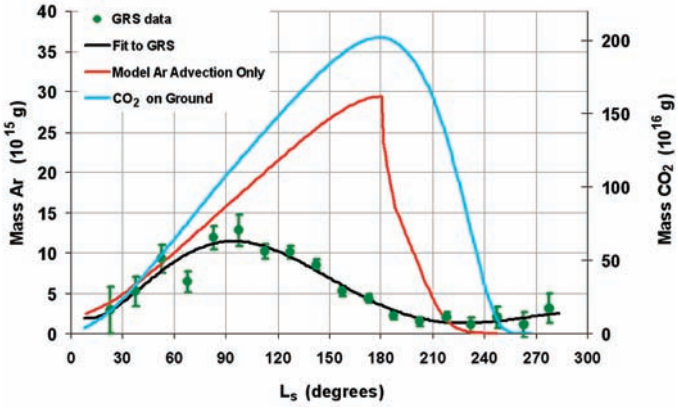


Fig. 2. The total mass of Ar over latitudes 75° to 90°S as measured by the GRS on Mars Odyssey is shown by filled green circles with error bars of the same significance as in Fig. 1. A sixth-order polynomial fit to the data is shown in black and is used for the computations. The red line is a model of the mass of Ar that would be expected above the southern polar region in the absence of eddy diffusion (mixing by transport owing to fluctuations in the mean winds). The blue line shows the mass of CO_2 frost on the ground. The bulk wind is driven by the condensation and subliming of this ground CO_2 frost deposit.

overturning cells or Hadley circulation could be involved, in principle (*12*), but such circulation is not predicted by most models to penetrate to the poles (*19*) and so we do not include it.

For computational convenience, we estimate df_{Ar}/dx by $\Delta f_{Ar}/\Delta x$, where Δf_{Ar} is the difference between the Ar mixing ratio measured by the GRS within the southern polar region and the seasonally adjusted Viking Lander 2 mixing ratio ($f_{GRS} - f_{Ar}$) at a distance Δx away. We define Δx as the distance Ar must be transported such that it is no longer observed in the southern polar footprint of the GRS and such that the ambient seasonal mixing ratio can be estimated. A good approximation is an arc length equivalent to 30° of latitude.

An estimate of K_x is found by the following rearrangement of Eq. 1:

$$K_x = \left(\frac{\Delta x}{\Delta f_{Ar}}\right)\left[v_x f_{Ar} - \left(\frac{F_{Ar}}{\rho}\right)\right] \qquad (2)$$

With our choice of signs, an equatorward Ar F_{Ar} is negative, and this is the case throughout most of winter and early spring (Fig. 3).

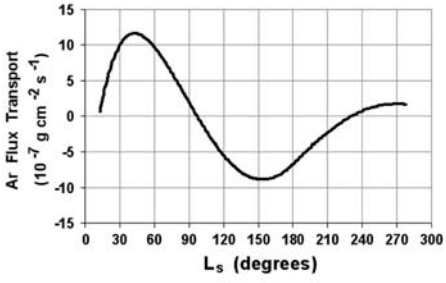


Fig. 3. Values for the Ar mass flux transport calculated from the fit to the GRS data are plotted versus L_s. Ar is carried into the polar region by the bulk wind throughout autumn and winter, but the mass of Ar transported to the polar region is diminished to zero in early winter by eddy meridional mixing toward the equator that offsets the condensation flow. Throughout late winter, the net transport of Ar is entirely toward the equator. By onset of spring, the subliming CO_2 from the polar region both dilutes and carries Ar toward the equator. In late spring, eddy meridional mixing begins to offset this condition, and the net mass transport is back into the polar region.

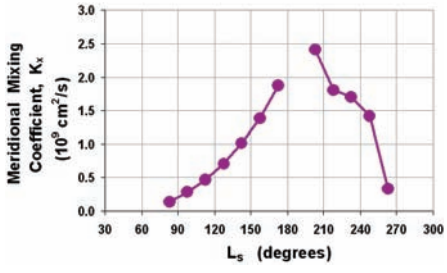


Fig. 4. Meridional eddy mixing coefficients for 15° increments in L_s throughout late autumn, winter, and spring calculated with the GRS data and the values for input shown in Table 1. The near coincidence of the GRS data (polynomial fit) and the advection-only Ar model during the period L_s 27° to 78° indicates no or little meridional eddy mixing (Fig. 2). In late autumn, meridional eddy mixing begins and its efficiency increases throughout winter. When the wind shifts direction at the onset of spring, the Ar mixing ratio drops to near zero and $\Delta x/\Delta f_{Ar}$ passes through zero, resulting in a large uncertainty in the mixing coefficient at that time. Thus, we do not show a value for eddy mixing at L_s 187°. Eddy transport is at a maximum in early southern spring and gradually diminishes as condensation flow begins to carry Ar toward the northern polar area.

Table 1. Relevant parameters from the computation of the eddy mixing coefficients with the use of Eqs. 1 and 2. The mixing ratios used for the ambient atmosphere can be obtained by adding the values in the fourth column to those from Fig. 1.

L_s (°)	v_x (cm s^{-1})	Wall area, 10^{14} (cm^2)	Δf_{Ar} GRS-VL2	F_{Ar}, 10^{-6} (g cm^{-2} s^{-1})	K_x, 10^8 (cm^2 s^{-1})
0–15	2.0	8.97			
15–30	3.0	8.20	−0.001	7.11	
30–45	4.2	8.00	0.009	11.3	
45–60	4.4	7.96	0.021	10.9	
60–75	4.7	7.95	0.031	7.99	
75–90	4.8	7.93	0.038	3.99	1.34
90–105	4.8	7.93	0.040	−0.196	2.88
105–120	5.0	7.92	0.038	−3.91	4.73
120–135	5.1	7.93	0.033	−6.76	7.07
135–150	4.8	7.95	0.025	−8.49	10.1
150–165	3.7	8.15	0.017	−8.82	13.8
165–180	1.6	8.72	0.008	−7.74	18.7
180–195	−3.1	9.97	0.001	−5.52	
195–210	−10.6	11.1	−0.004	−3.32	24.1
210–225	−20.2	12.0	−0.008	−1.29	18.0
225–240	−26.3	12.8	−0.009	0.122	17.0
240–255	−15.2	12.0	−0.008	1.12	14.1
255–270	−0.03	13.5	−0.006	1.59	3.26
270–285		13.6	−0.004	1.62	

The GRS measurements of Ar are made by summing gamma ray counts from 75° to 90°S over 15° of L_s. To remove discontinuities from the computations, we fitted the GRS data with a sixth-order polynomial curve (Fig. 2). To compute the F_{Ar}, we take the time rate of change of Ar mass, assuming it is vertically mixed (*20*, *21*), as it passes through an imaginary cylinder surrounding the polar region. The cylinder corresponds to the 75°S latitude circle and a cylindrical wall, the area of which varies seasonally because of the change in atmospheric scale height H (*22*). Wall areas are shown in Table 1. The time used is the number of seconds appropriate for the exact interval. We compute the value of K_x for every 1° increment of L_s.

The GRS Ar mass mixing ratios are greater than the nominal Viking Lander 2 value throughout autumn and winter but less during the entire spring (Fig. 1). A peak in Ar has been reached by the onset of winter. From that time on, the Ar mixing ratio decreases even though CO_2 frost continues to accumulate on the south polar cap (Fig. 2). The Ar mixing ratio does not continue to increase throughout winter, because equatorward mixing removes Ar from the polar area.

In spring, when sunlight hits the polar cap, a rapid increase in atmospheric CO_2 occurs as CO_2 frost sublimes. This creates a substantial dilution of Ar as CO_2 is added to the polar atmosphere. At this time the wind shifts in direction, increases speed, and blows equatorward, out of the high-latitude region. The Ar mixing ratio does not decrease to an unmeasurably low quantity (Fig. 2), indicating that Ar is now mixed down the gradient toward the southern polar region against the southerly wind. The transport of Ar back to the southern polar region becomes positive in rapid response to the CO_2 dilution and wind. The process is driven by the gradient in the mixing ratio and fluctuations in the mean winds (eddy mixing). This is implicit in the positive slope of the Ar flux transport curve between L_s 160° and 270° (Fig. 3).

The eddy mixing coefficients, K_x, that were derived from the data (Fig. 4) change with season. Because the GRS measurements track the Ar predicted in our advection-only model until late autumn, the dominant mechanism of Ar transport must be condensation flow, with negligible meridional eddy mixing during this period. However, the GRS data depart from the model in late autumn (Fig. 2), indicating that the gradient in Ar mixing ratio has increased to the point where eddy mixing begins. This change in the nature of the polar atmosphere is consistent with analysis of the Mars global surveyor thermal emission spectrometer

(TES) data set for the same time periods. In the TES analysis, vertical temperature profiles have been inverted to model potential vorticity and zonal wind fields over both the southern and northern polar winter seasons (*23*). The northern annular structure is retained throughout winter, but the southern zonal structure is chaotic and exhibits fluctuations at the end of autumn (*24*). Fluctuations are consistent with what might be expected with local weather variations. Meridional mixing increases throughout winter and apparently reaches a maximum in early spring after the initial addition of new CO_2 into the polar atmosphere when the polar CO_2 cap begins to sublime. By the end of spring, the Ar mixing ratio has been stabilized and the eddy mixing diminishes (Fig. 4).

The GRS data have measured the vertically integrated transport of Ar in and out of the southern polar atmosphere. In addition, the measurements have permitted derivation of mixing characteristics of Mars' southern polar atmosphere that are different from those observed in Earth's stratosphere. The active and increasing equatorward mixing throughout Mars' winter is in contrast to the Earth's stratosphere, in which, at similar pressures and during the same season, air over the southern cap is isolated by the southern vortex for the entire winter (*25*). On Earth, this isolation permits the action of unusual chemistry that leads to the spring "ozone hole," which lasts about a month as the southern vortex dissipates. On Mars, it is evident that by late autumn no strong containment of Ar exists. In autumn and winter, the substantial poleward flow of CO_2 that condenses onto the polar cap brings with it a substantial amount of Ar, but in winter the equatorward mixing of Ar toward lower Ar concentration present in "average" air at mid-latitudes effectively clears excess Ar out of the polar region. In spring, the CO_2 wind transports Ar out of the polar region to lower latitudes, and atmospheric pressure returns to its southern summer value. During this time, higher mixing ratios of Ar in the average atmosphere at mid-latitudes create gradients that cause eddy mixing of Ar back into the southern polar region. The magnitude of the Ar fluxes derived from the GRS measurements provides real data that are important for understanding meridional mixing at southern polar latitudes and for constraining MGCMs that attempt to predict martian meteorology.

References and Notes

1. L_s, the areocentric longitude of the Sun as measured in a Mars-centered fixed coordinate system, is associated with the seasons on Mars. For the southern hemisphere, L_s 0°, 90°, 180°, and 270° correspond to the onset of autumn, winter, spring, and summer, respectively.
2. H. H. Kieffer *et al.*, *Science* **193**, 780 (1976).
3. H. H. Kieffer, T. Z. Martin, A. R. Peterfreund, B. M. Jakosky, E. E. Miner, F. D. Palluconi, *J. Geophys. Res.* **82**, 4249 (1977).
4. S. L. Hess, *J. Geophys. Res.* **84**, 2969 (1979).
5. W. V. Boynton *et al.*, *Space Sci. Rev.* **110**, 37 (2004).
6. W. D. Feldman *et al.*, *J. Geophys. Res.* **108**, 10.1029/2003JE002101 (2003).
7. M. L. Litvak *et al.*, *Solar Sys. Res.* **38**, 167 (2004).
8. T. H. Prettyman, R. C. Wiens, J. R. Murphy, J. M. Reisner, W. C. Feldman, *35th Lunar Planet. Sci. Conf. No. 1878* (2004).
9. A. E. Metzger, E. L. Haines, *J. Geophys. Res.* **95**, 14695 (1990).
10. The GRS footprint is somewhat smaller than the south polar region analyzed with 50% of the signal coming from a region of radius 4° of arc (about 240 km). Data are collected when the spacecraft is within 15° of the pole, so we are actually sampling down to about 70° S latitude.
11. R. M. Haberle, personal communication.
12. J. B. Pollack, R. M. Haberle, J. Schaeffer, H. Lee, *J. Geophys. Res.* **95**, 1447 (1990).
13. A martian solar day has a mean period of 24 hours 39 min 35.244 s and is referred to as a sol to distinguish this from a roughly 3% shorter solar day on Earth.
14. T. Owen *et al.*, *J. Geophys. Res.* **82**, 4635 (1977).
15. J. W. Chamberlain, D. M. Hunten, *Theory of Planetary Atmospheres* (Academic Press, New York, 1987).
16. T. H. Prettyman *et al.*, *J. Geophys. Res.* **109**, EO5001, 10.1029/2003JE002139 (2004).
17. N. Kelly *et al.*, *6th International Conference on Mars No. 3244* (2003).
18. The north-south wind speed can be defined as follows:

$$v_x(\mathrm{cm/s}) = \frac{\mathrm{CO_2\ mass\ transport\ (g/s)}}{\mathrm{Viking\ Lander\ 2\ mean\ density\ (g/cm^3)} \times \mathrm{wall\ area\ (cm^2)}}$$

19. R. J. Wilson, *Geophys. Res. Lett.* **24**, 123 (1997).
20. A. Colaprete, R. M. Haberle, O. B. Toon, *J. Geophys. Res.* **108**, 10.1029/2003JE002053 (2003).
21. D. Banfield, B. J. Conrath, M. D. Smith, P. R. Christesen, J. R. Wilson, *Icarus* **161**, 319 (2003).
22. The atmospheric scale height, H, is the vertical distance over which the atmospheric pressure drops by a factor of e (2.718…). The scale height undergoes an increase in spring from the average autumn and winter value of 7.5 km to nearly 12.5 km by the start of summer, causing an increase in the wall area that is specified in Table 1.
23. T. H. McConnochie, B. J. Conrath, P. J. Gierasch, D. Banfield, M. D. Smith, *6th International Conference on Mars No. 3249* (2003).
24. T. H. McConnochie, B. J. Conrath, D. Banfield, P. J. Gierasch, M. D. Smith, *AAS Division for Planetary Sciences* (American Astronomical Society, Washington, DC, 2003), pp. 14–17.
25. M. R. Schoeberl, D. L. Hartmann, *Science* **251**, 46 (1991).

26. We thank all the members of the GRS team for the years of dedicated efforts to make these and other measurements. We are especially grateful to K. Crombie, D. Drake, L. Evans, C. Fellows, O. Gasnault, D. Hamara, N. Kelly, R. Starr, and J. Trombka, who contributed to the gamma ray analysis. We also thank R. Haberle, who made substantial contributions to this work with the support of run 2002.17 of the NASA Ames MGCM and with many helpful discussions. The GRS and its scientists are supported through NASA contract no. 1228726.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1098496/DC1
SOM Text

30 March 2004; accepted 24 September 2004
Published online 7 October 2004;
10.1126/science.1098496
Include this information when citing this paper.

Very Low Gene Duplication Rate in the Yeast Genome

Li-zhi Gao and Hideki Innan[1]*

The gene duplication rate in the yeast genome is estimated without assuming the molecular clock model to be ~0.01 to 0.06 per gene per billion years; this rate is two orders of magnitude lower than a previous estimate based on the molecular clock model. This difference is explained by extensive concerted evolution via gene conversion between duplicated genes, which violates the assumption of the molecular clock in the analyses of duplicated genes. The average length of the period of concerted evolution and the gene conversion rate are estimated to be ~25 million years and ~28 times the mutation rate, respectively.

Gene duplication is considered to be an important mechanism for the generation of genomic novelty (*1*, *2*). A crucial question is the rate at which gene duplication occurs. Lynch and Conery (*3*) estimated this rate with the use of complete genome sequences of three model eukaryote species, including yeast. Because comparative genomic data were not available at that time, the molecular clock model (*4*) was assumed to estimate the rate from a single genomic sequence. Their estimates of the gene duplication rate per gene were surprisingly high—roughly on the order of one per 100 million years. However, although the molecular clock model should be reasonably accurate when applied to sequence data between species (*2*), it is not clear whether the molecular clock works for the divergence between duplicated genes, as gene conversion may homogenize interlocus variation. This nonindependent evolution of copy members in a multigene family is known as concerted evolution (*5*–*7*). If concerted evolution is a common phenomenon, it is expected that a molecular clock–based estimate of the gene duplication rate should be

[1]Human Genetics Center, School of Public Health, University of Texas Health Science Center, 1200 Hermann Pressler, Houston, TX 77030, USA.

*To whom correspondence should be addressed. E-mail: hideki.innan@uth.tmc.edu

inflated (*8*). Here, we report a method for estimating the gene duplication rate without the molecular clock assumption, and we derive a much lower rate of gene duplication.

The complete genome sequences of *Saccharomyces cerevisiae* (*9*) and six of its relatives (*10*, *11*) were used to estimate the gene duplication rate without assuming the molecular clock model. On the basis of these genomic data, a highly reliable species tree is provided (*12*) (Fig. 1A). *Ks*, the synonymous nucleotide divergence, is used to measure the time of the speciation events, which are denoted by T_1, T_2, T_3, T_4, T_5, and T_6 in chronological order. In this study, we "mapped" the timings of recent gene duplication events in the intervals between the nodes on the species tree, and from this we derived the rate of duplication.

In the *S. cerevisiae* genome, we identified 68 complete duplicated genes (i.e., two-copy gene families) with $Ks < 1.05$ (*13*). Duplicated copies for most of them are located on different chromosomes. Because $Ks = 1.05$ approximately corresponds to the divergence between *S. cerevisiae* and *S. bayanus*, most of these duplications should have occurred after T_4 under the prediction of the molecular clock model. The two duplicated genes for each pair were randomly denoted by X and Y (Fig. 1B). The two adjacent genes (A and B for X; C and D for Y) were used to examine whether the orthologs of X and Y exist in the whole-genome draft sequences of the six relatives of *S. cerevisiae* (*13*). The results appear in table S1 (a portion of the results is shown in Table 1).

Each entry in Table 1 (table S1) consists of candidates of the orthologs of X and Y and the number of BLAST hits of the focal duplicates (X and Y) for which the orthology was not successfully estimated. For example, the ninth pair of duplicated genes (X: YBR181C; Y: YPL090C) indicates that the focal pair has two BLAST hits in *S. kudriavzevii*, one with A and B and the other with C and D (AXB/CYD/0). Therefore, the synteny around X and Y is assumed to be conserved in *S. kudriavzevii*, which suggests that the duplication event is older than T_3, because independent duplications inserted into the same gene interval should be extremely unlikely. Support is provided by the result for *S. mikatae* (AX-/CYD/1), where orthologous parts of both of the focal gene pairs are found (with an extra BLAST hit for the duplicates), although we were not able to identify the ortholog of X in *S. paradoxus* (---/CYD/1). For *S. bayanus*, *S. castellii*, and *S. kluyveri*, it was not possible to obtain sufficient evidence that the synteny around X and Y is conserved. In this way, we can estimate T_m, the minimum age of the duplication event (in this case, T_3).

The gene duplication rate can be directly estimated from T_m without assuming the molecular clock model. Only one duplication event must have occurred between T_0 and T_1. For this pair (first gene pair, X: YHR053C; Y: YHR055C), the ancestral state at T_1 was inferred to be AXB and a tandem gene duplication event created AXYB in *S. cerevisiae*. Although there are several pairs with $T_m = T_0$, it is not clear whether they were created after T_1. Therefore, the number of gene duplication events between T_0 and T_1 may be from one to five. Given that the average *Ks* between *S. cerevisiae* and *S. paradoxus* is ~0.36, the ratio of the gene duplication rate per genome to the synonymous substitution rate per site is estimated to be (1 to 5)/0.18 = 5.6 to 28. If the synonymous substitution rate per site per year is assumed to be 8.1×10^{-9} (*3*) and the number of single-copy genes in the yeast genome is ~3500 (*9*), the gene duplication rate is ~0.01 to 0.06 per billion years.

We also used our data to estimate the gene duplication rate with the molecular clock–based method, because our estimate cannot be directly comparable to the estimate reported by Lynch and Conery (*3*). Because there are five gene pairs with $Ks < 0.01$, the gene duplication rate is estimated to be 2.3 per billion years, two orders of magnitude larger than our nonclock-based estimate. This difference is highly significant. If we suppose that our nonclock-based estimate is correct, the expected number of duplicated genes with $Ks < 0.01$ is less than 0.14. Then, the probability of observing five or more gene duplication events is less than 4×10^{-7}. Note that by "gene duplication rate" we mean the rate at which a duplicated gene is created by mutation and becomes fixed in the population. The fixation probability of duplicated genes should be largely affected by natural selection (*14*).

The difference between the molecular clock–based and nonclock-based estimates could be explained by extensive concerted evolution via gene conversion. With gene conversion, many old duplicated genes can appear as if they are young. Note that the molecular clock predicts that duplicated genes with low nucleotide divergence are "young." This fact causes the inflation of a molecular clock–based estimate of gene

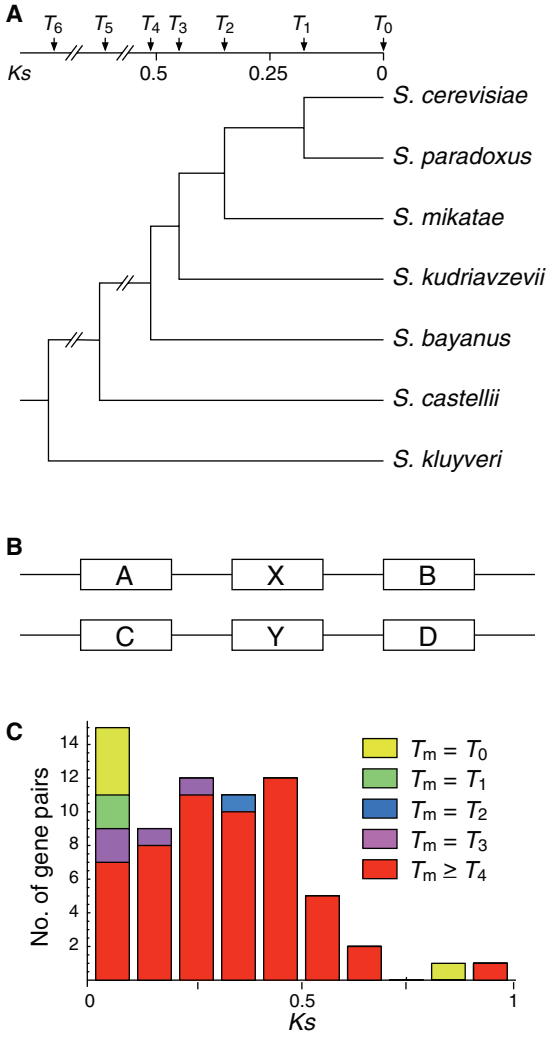


Fig. 1. (**A**) Species tree for the seven yeast species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, and *S. kluyveri*) estimated from 106 widely distributed orthologous genes (*12*). Using these 106 genes, the average synonymous nucleotide divergence (*Ks*) was estimated by the PAML package (*27*). (**B**) Strategy for identifying the orthologs of a pair of duplicated genes, X and Y. (**C**) Frequency distribution of *Ks* between gene pairs in *S. cerevisiae*.

duplication rate, because it depends on the number of duplicated genes that look young. Evidence for extensive gene conversion is provided by two estimated gene trees (Fig. 2). The two duplicates in each species are more closely related to each other than to the orthologs of other species with high bootstrap support. Therefore, we tested the applicability of the molecular clock model to duplicated genes. The frequency distribution of *Ks* between two duplicated genes for each T_m is shown in Fig. 1C. It is obvious that most pairs (55/68 = 81%) have $T_m \geq T_4$, whereas the molecular clock predicts that most duplicated genes should be younger than T_4. We used 66 gene pairs with $Ks + 1.96 \times SD < 1.05$, where SD is the standard deviation of *Ks*. For these pairs, because the probability that the real age predates T_4 may be <0.025 under the null hypothesis, the expected number of gene pairs with $T_m \geq T_4$ should be smaller than $66 \times 0.025 = 1.65$. Even with this conservative value, our observation of 55 pairs with $T_m \geq T_4$ clearly rejects the null hypothesis ($P = 10^{-76}$). Figure 1C indicates that duplicated genes created before T_4 may have very small values of *Ks*; this suggests that it is very difficult to estimate the age of a duplication event from *Ks*, although some gene pairs with low *Ks* may be young. Another piece of evidence for extensive gene conversion is that many of the analyzed gene pairs should have been created by the whole-genome duplication event that occurred 100 to 150 million years ago (*15*). Of the 68 gene pairs, we found that 37 were in the genome duplication blocks recently identified by Kellis *et al*. (*16*). Some of them have very low values of *Ks*, in contrast to the expectation of $Ks \gg 1$ under the molecular clock hypothesis.

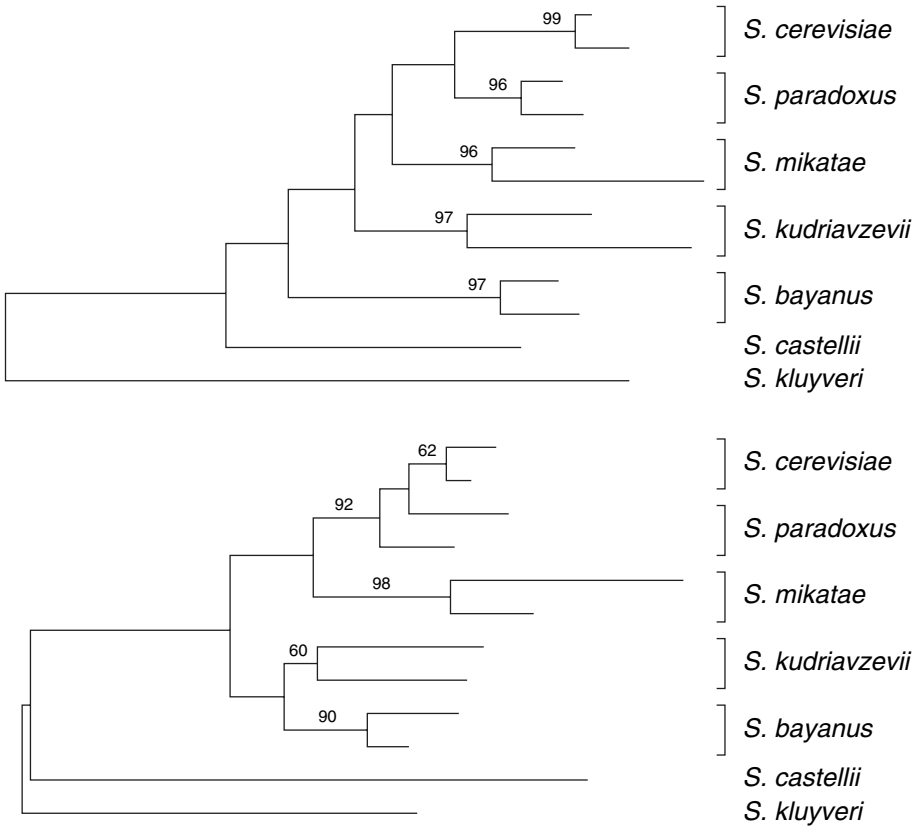


Fig. 2. Evidence for extensive concerted evolution shown in neighbor-joining (NJ) gene trees. (**A**) Gene tree for the orthologs of YGL135W and YPL220W (the sixth gene pair in Table 1). (**B**) Gene tree for the orthologs of YDL136W and YDL191W (the eighth gene pair in Table 1).

Many old duplicate genes that look young also cause the inflation of a molecular clock–based estimate of the gene loss rate (*8*). The gene loss rate can be estimated by the distribution of the age of duplicated genes in a genome, so that an estimate is very sensitive to estimates of ages. If we underestimate ages [e.g., by assuming the molecular clock (*3*)], the distribution is skewed toward younger genes, causing an overestimation of the gene loss rate. Although we were not able to estimate the gene loss rate from our data, the gene loss rate may not be much higher than the gene duplication rate. We have estimated the gene duplication rate from duplicated genes with $T_m \leq T_1$ and $T_m \leq T_2$, and these estimates (0.007 to 0.05 per billion years and

Table 1. List of the studied duplicate gene pairs of *S. cerevisiae* and their orthology in its six relatives. See table S1 for the full list of the investigated gene pairs.

	X	Y	*Ks*	*S. paradoxus*	*S. mikatae*	*S. kudriavzevii*	*S. bayanus*	*S. castellii*	*S. kluyveri*	T_m*
1†	YHR053C	YHR055C	0	AXB/0	AXB/0	AX-/0	AXB/0	---/0	---/0	0
2	YCL066W	YCR040W	0	AXB/-YD/2	---/CYD/2	-XB/CYD×2/1	AXB/-YD/1	A-B/---/0	---/---/0	4
3	YNL019C	YNL033W	0	AXB/-YD/1	---/-YD/0	---/CYD/0	AX-/CYD/0	---/---/0	---/---/0	1
4	YAR064W	YHR213W-B	0	---/-YD/3	---/---/0	---/---/1	---/---/0	---/---/0	---/---/0	0
5	YFL061W	YNL335W	0.0053	---/-YD/0	---/---/1	---/---/0	---/---/0	---/---/0	---/---/0	0
6	YGL135W	YPL220W	0.0209	AXB/CYD/0	AXB/CYD/0	-XB/-YD/0	AXB/CY-/0	A-B/CYD/0	AX-/---/0	4
7	YOR390W	YPL279C	0.0209	AXB/---/0	A-B/---/1	-XB/---/1	---/---/1	---/---/0	---/---/0	0
8	YDL136W	YDL191W	0.0265	AXB/CYD/0	-XB/CY-/0	AXB/CY-/0	AXB/CYD/0	A-B/CYD/0	---/CYD/0	4
9	YBR181C	YPL090C	0.0298	---/CYD/1	AX-/CYD/1	AXB/CYD/0	AXB/---/1	---/CYD/1	---/---/1	3
10	YNL018C	YNL034W	0.0329	AXB/CY-/1	---/CY-,-YD/0	---/CYD/0	---/CYD/1	---/---/0	---/---/0	1
11	YBR031W	YDR012W	0.0354	AXB/CYD/0	AXB/CY-, -YD/0	AXD/CYB/0	AXD/CYB/0	AXD/---/0	AX-/---/0	4
12	YHR141C	YNL162W	0.0677	AXB/CYD/0	-XB/CYD/0	AXB/CYD/0	AXB/CYD/0	AXB/-YD/0	---/---/1	5
13	YHR203C	YJR145C	0.0699	AXB/CYD/0	AXB/CYD/0	AXB/---/2	AXB/CY-/0	---/---/2	---/---/1	4
14	YER074W	YIL069C	0.0699	AXB/CYD/0	AXB/CYD/0	AXB/CYD/0	AXB/CYD/0	---/CY-/1	---/---/1	4
15	YBL072C	YER102W	0.0947	AXB/CYD/0	AX-/CYD/0	AX-/CYD/0	AX-/---/1	---/CY-/1	---/---/1	3
16	YBR009C	YNL030W	0.1339	AXB/CYD/0	AXB/CYD/0	---/---/2	-XB/CYD/0	-XB/CYD×2/0	---/CY-/1	5
17	YHL001W	YKL006W	0.1343	AXB/CYD/0	AX-/-YD/1	AXB/CYD/0	AXB/CY-/0	-XB/CYD/0	---/---/1	5
18	YIL018W	YFR031C-A	0.1412	AXB/CYD/0	AX-/CYD/0	AXB/CYD/0	AXB/CY-/0	-XB/---/1	---/CY-/0	4
19	YGR085C	YPR102C	0.1445	AXB/CYD/0	AX-,-XB/CYD/0	AXB/CYD/0	AXB/CYD/0	---/---/1	---/---/1	4
20	YHL033C	YLL045C	0.1457	AXB/CYD/0	AXB/CY-/0	AXB/CYD/0	AXB/CYD/0	---/---/1	---/---/0	4

*Only the subscript of T_m is shown. †Tandem duplicated genes for which the gene order of X, Y, and markers is given by AXYB in *S. cerevisiae*.

0.005 to 0.04 per billion years, respectively) are similar to the estimate from genes with $T_m = T_0$, which is not expected if the gene loss rate is much higher than the gene duplication rate.

The data also allow us to quantify the duration of concerted evolution and the level of gene conversion. The question of the duration of concerted evolution is addressed according to the recent theoretical result of Teshima and Innan (*8*), who showed that the period of concerted evolution approximately follows an exponential distribution with parameter $1/\tau$, where τ is the expected length of concerted evolution. The probability (f) that the duration of concerted evolution from a certain time point (t_s) exceeds another time point (t_e) is given by $\exp[-(t_e - t_s)/\tau]$. To estimate τ assuming a constant τ for all gene pairs, we considered two time points, T_4 and T_1, on the species tree (Fig. 1A). We focused on the 51 gene pairs for which concerted evolution was likely occurring at T_4. For each of these 51 gene pairs, we considered whether concerted evolution was still going on at T_1 by comparing *Ks* among four gene sequences, two from *S. cerevisiae* and two from *S. paradoxus*. We found smaller values of *Ks* between the paralogs within species than between orthologs for nine gene pairs that were considered to be under concerted evolution at T_1. We could then estimate $f = 9/51$, from which an estimate of $\tau = 0.2$ was obtained by solving $\exp(-0.35/\tau) = f$, where 0.35 is the time between T_1 and T_4 measured in units of $1/Ks$. Assuming the synonymous substitution rate $Ks = 8.1 \times 10^{-9}$, the estimate yields $\tau = 25$ million years (*13*).

The rate of gene conversion is one of the important factors in determining the period of concerted evolution. The gene conversion rate can be directly estimated from the nucleotide divergence between gene pairs currently under concerted evolution. Because it was not possible to determine such gene pairs, we used the nine gene pairs that are likely under concerted evolution at T_1 as a proxy, for which the average $d = 0.036$. The expectation of d is given by μ/c, where μ is the mutation rate per site and c is the gene conversion rate per site (*17*, *18*); hence, we estimate that the gene conversion rate is ~28 times the mutation rate, assuming that c is constant for all duplicated genes (*13*). This is within the range of estimates (10 to 100) in *Drosophila* duplicated genes (*18*).

Our demonstration of extensive concerted evolution via gene conversion on a genome scale is consistent with molecular genetic studies showing frequent interlocus gene conversion in yeast (*19*). Although yeast is a model species for studying gene conversion, there is no reason to believe that the effect of gene conversion in duplicated genes is negligible in other organisms. Increasing evidence for gene conversion (interlocus as well as intralocus) is also available in higher eukaryotes, such as humans (*20–23*), *Drosophila* (*18*, *24*, *25*), and other species (*26*).

References and Notes

1. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).
2. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
3. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
4. E. Zuckerkandl, L. Pauling, in *Evolving Genes and Proteins*, V. Bryson, H. J. Vogel, Eds. (Academic Press, New York, 1965), pp. 97–166.
5. T. Ohta, *Evolution and Variation of Multigene Families* (Springer-Verlag, Berlin, 1980).
6. G. Dover, *Nature* **299**, 111 (1982).
7. J. F. Elder Jr., B. J. Turner, *Q. Rev. Biol.* **70**, 297 (1995).
8. K. M. Teshima, H. Innan, *Genetics* **166**, 1553 (2004).
9. A. Gofieau *et al.*, *Science* **274**, 546 (1996).
10. P. Cliften *et al.*, *Science* **301**, 71 (2003).
11. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, *Nature* **423**, 241 (2003).
12. A. Rokas, B. L. Williams, N. King, S. B. Carroll, *Nature* **425**, 798 (2003).
13. See supporting data on *Science* Online.
14. B. Walsh, *Genetica* **118**, 279 (2003).
15. K. Wolfe, D. Shields, *Nature* **387**, 708 (1997).
16. M. Kellis, B. W. Birren, E. S. Lander, *Nature* **428**, 617 (2004).
17. T. Ohta, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3251 (1982).
18. H. Innan, *Genetics* **163**, 803 (2003).
19. T. D. Petes, C. W. Hill, *Annu. Rev. Genet.* **22**, 147 (1988).
20. S. Rozen *et al.*, *Nature* **423**, 873 (2003).
21. H. Innan, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8793 (2003).
22. M. E. Hurles, D. Willey, L. Matthews, S. S. Hussain, *Genome Biol.* **5** (8), 10.1186/gb-2004-5-8-r55 (2004).
23. A. J. Jeffreys, C. A. May, *Nature Genet.* **36**, 151 (2004).
24. C. H. Langley, B. P. Lazzaro, W. Phillips, E. Heikkinen, J. M. Braverman, *Genetics* **156**, 1837 (2000).
25. W. Wang, K. Thornton, J. J. Emerson, M. Long, *Genetics* **166**, 1783 (2004).
26. G. Marais, *Trends Genet.* **19**, 330 (2003).
27. Z. Yang, R. Nielsen, *Mol. Biol. Evol.* **17**, 32 (2000).

28. We thank S. A. Barton, Y.-X. Fu, Z. Gu, Y. Jin, M. Long, M. Nordborg, T. Ohta, N. Rosenberg, K. M. Teshima, K. Thornton, three anonymous referees for comments, and A. Rokas for data. H.I. is supported by a grant from the University of Texas.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1367/DC1
Materials and Methods
SOM Text
Table S1
Figs. S1 and S2
References

28 June 2004; accepted 4 October 2004

Regulated Fast Nucleocytoplasmic Shuttling Observed by Reversible Protein Highlighting

Ryoko Ando, Hideaki Mizuno, Atsushi Miyawaki*

The observation of the regulation of fast protein dynamics in a cellular context requires the development of reliable technologies. Here, a signal regulation cascade reliant on the stimulus-dependent acceleration of the bidirectional flow of mitogen-activated protein kinase (extracellular signal-regulated kinase) across the nuclear envelope was visualized by reversible protein highlighting. Light-induced conversion between the bright and dark states of a monomeric fluorescent protein engineered from a novel coral protein was employed. Because of its photochromic properties, the protein could be highlighted, erased, and highlighted again in a nondestructive manner, allowing direct observation of regulated fast nucleocytoplasmic shuttling of key signaling molecules.

Fluorescent proteins isolated from cnidaria have been used to generate fusion tags that act as fluorescent reporters for studying protein dynamics (*1–3*). However, tracking protein behavior is difficult when the tagged protein is evenly distributed under steady-state conditions. Fluorescence recovery after photobleaching (FRAP) and optical highlighting of fluorescent proteins can provide insights into the diffusive or directed movement of proteins and track rapid protein behavior (*4–7*), but protein movement is regulated by many different factors and may be altered by changes in the cellular state. Because photobleaching, photoactivation, and photoconversion with the available markers are often irreversible or complex, the identification of genetically encoded fluorescent proteins that can be marked reversibly for repeated measurements of protein behavior has long been pursued.

We prepared a cDNA library from Pectiniidae, a species of coral that emits faint fluorescence upon irradiation with ultraviolet light. Approximately 300,000 bacterial colonies containing individual cDNA clones were screened for fluorescence (*8*). A single clone (22G) encoding a greenish fluorescent protein with highest homology to mcavGFP

Laboratory for Cell Function and Dynamics, Advanced Technology Development Group, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako-city, Saitama, 351-0198, Japan.

*To whom correspondence should be addressed. E-mail: matsushi@brain.riken.jp

(*9*) (74.6%) was identified. Recombinant 22G protein was expressed in *E. coli* and purified, and its molecular mass determined to be 102 kD by analytical equilibrium ultracentrifugation analysis. This was 3.5 times as large as that predicted from the primary structure of the protein (29.2 kD), which suggests that 22G forms an oligomeric complex (*2*). We engineered a monomeric version of 22G (22Gm3) by introducing both rational and random mutations (*10*) (fig. S1). The molecular mass of 22Gm3 was measured to be 28.8 kD.

The absorption spectrum of 22Gm3 at pH 7.4 displayed a major peak at 503 nm (ε = 95,000 $M^{-1}\cdot cm^{-1}$) and a minor peak at 390 nm. The amplitude of the 503-nm peak decreased with decreasing pH, whereas that of the 390-nm peak increased, exhibiting an isosbestic point at 428 nm (Fig. 1A). The 390- and 503-nm peaks correspond respectively to the neutral and ionized states of the phenolic hydroxyl of the chromophore (*1*). The apparent pK_a was 5. Excitation and emission spectra of 22Gm3 were analyzed (Fig. 1C). The neutral form was nonfluorescent, but the ionized form was highly fluorescent, with an emission maximum of 518 nm. The fluorescence quantum yield (Φ_{FL}) was 0.85.

We found 22Gm3 to possess a distinctive photosensitivity. Strong excitation at around 490 nm appeared to bleach 22Gm3 more efficiently than other fluorescent proteins, and the bleached protein regained its green fluorescence completely with minimal irradiation at around 400 nm. We intermittently measured the absorbance of a solution containing 22Gm3 at pH 7.4 in a cuvette during continuous illumination at 490 ± 10 nm with a 75-W xenon lamp (Fig. 1B). After a 40-min incubation, nearly all the protein molecules had been converted into the neutral, nonfluorescent state. Illumination at 400 ± 7.5 nm for several minutes reversed the protein to its original fluorescent state (Fig. 1B). Thus, 22Gm3 has photochromic behavior, because its fluorescence can be switched on and off by using two different wavelengths of light. This photochromism is not pH sensitive, because the neutral state generated by acidification was unaffected by illumination with 400-nm light (fig. S2). Based on the capacity of its fluorescence to vanish and reappear, 22Gm3 was renamed "Dronpa," after "dron," a ninja term for vanishing, and "pa," which stands for photoactivation.

The kinetics of the photochromic behavior of Dronpa were explored in cells by microscopy (*8*). When Dronpa was transfected into HeLa cells, green fluorescence was uniformly distributed in both cytosolic and nuclear compartments (Fig. 1D). Fluorescence was monitored in fixed cell samples subjected to strong irradiation at 400 or 490 nm (Fig. 1E). Photoactivation of Dronpa required much less photon energy than did

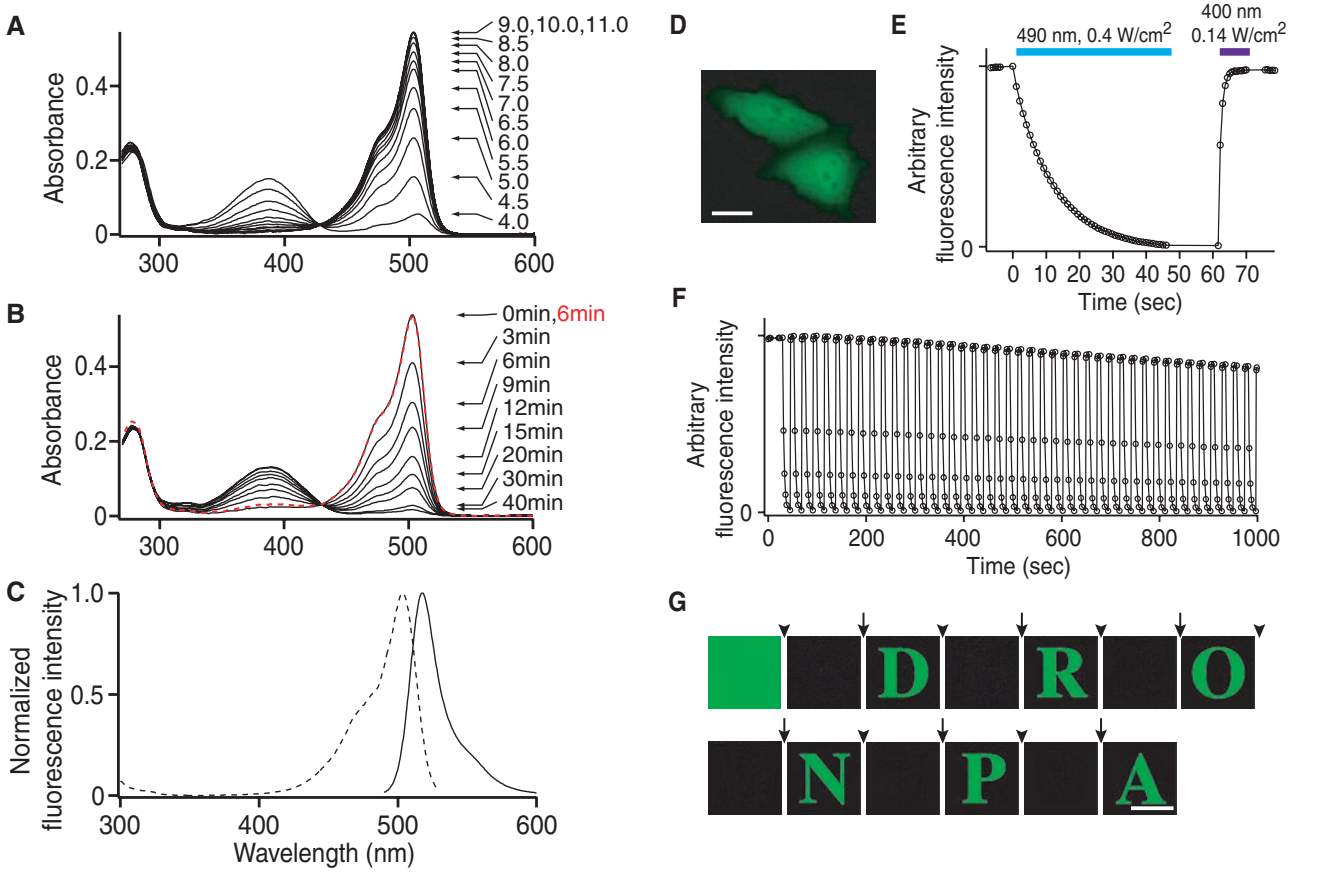


Fig. 1. Photochromic properties of Dronpa (22Gm3). (**A**) The pH dependence of 22Gm3 absorbance. (**B**) Irradiation-dependent changes in 22Gm3 absorbance. Absorbance spectra obtained during irradiation at 490 nm (black) and after irradiation at 400 nm (red) are displayed. (**C**) Normalized excitation (dotted line) and emission (solid line) spectra of 22Gm3. (**D**) Fluorescence micrograph of HeLa cells expressing Dronpa. Scale bar, 20 μm. (**E**) Time course of intensity of Dronpa fluorescence in a fixed HeLa cell. Green fluorescence was monitored by using a 490DF20 excitation filter overlaid with a 0.3% transmittance neutral density (ND) filter, a 505DRLPXR dichroic mirror, and a 535DF25 emission filter. During the marked intervals, the cell was continuously illuminated through a 30% transmittance ND filter at 490 nm (490DF20; 0.40 W/cm^2) or 400 nm (400DF15; 0.14 W/cm^2) to induce photobleaching or photoactivation, respectively. (**F**) On/off cycles of Dronpa fluorescence observed using the same cell and filter set as in (E), except that excitation intensities were 1.3 W/cm^2 (490DF20) and 0.47 W/cm^2 (400DF15) for photobleaching and photoactivation, respectively. After 100 cycles, fluorescence intensity of the fully activated sample was 75% of the original level. (**G**) Six letters written successively on the same field of a cover slip overlaid with immobilized Dronpa protein. The time points of erasure at 488 nm and writing at 405 nm are indicated by arrowheads and arrows, respectively. Scale bar, 300 μm.

photobleaching, with respective quantum yields (Φ_{PA} and Φ_{PB}) of 0.37 and 0.00032. Fluorescence could be switched on and off repeatedly (Fig. 1F). The off state was thermally stable; <1% of the original intensity was recovered after incubation at 20°C for 10 min in the dark. Using a laser-scanning confocal microscope, we were also able to inscribe six letters on a cover slip with immobilized Dronpa protein (*8*) (Fig. 1G), showing that Dronpa is an information storage medium with the ability to record, erase, or read information nondestructively (*11*, *12*). Comparisons between Dronpa and photoactivatable (PA)-GFP (*5*) are shown in table S1. The relatively fast bleaching rate of Dronpa at 488 nm (Φ_{PB} = 0.00032) may be a drawback because a limited number of images can be acquired after photoactivation. However, because of the high Φ_{FL} (0.85), multiple bright images could be obtained when the excitation light was reasonably attenuated.

Although photochromism has been reported for the yellow-emitting variants (YFPs) of *Aequorea* GFP at the single-molecular level (*13*), they have not demonstrated appreciable photochromism when measured in bulk. After a single round of photobleaching with illumination at 490 nm, only a small fraction (10 to 40%) of the original intensity could be restored (*14*), and recovery did not necessarily require illumination with 400-nm light, indicating the thermal instability of the off state of YFP.

The perfect photochromic properties of Dronpa suggested that it would be a useful tool to analyze the heretofore unexplored regulation of fast protein dynamics. The mitogen-activated protein kinase [extracellular signal-regulated kinase (ERK)] signaling cascade transduces a variety of extracellular signals at multiple levels in the cell (*15*, *16*). Upon stimulation, ERK detaches from its cytosolic anchors and translocates into the nucleus, where it regulates gene expression by phosphorylating transcription factors. ERK is then returned to the cytoplasm through nuclear export mechanisms that are sensitive to leptomycin B, a specific inhibitor of nuclear export signal (NES)–mediated nuclear export. Previous studies with immunocytochemistry or GFP-fused ERK (*17*, *18*) have examined only nuclear accumulation upon leptomycin B treatment or the steady-state distribution of ERK between the nucleus and cytoplasm, which should be dictated by the relative strengths of the nuclear import and export flows.

To measure directly both nuclear influx and efflux of ERK, we tagged ERK1 with Dronpa (ERK1-Dronpa) and examined its behavior in COS7 cells upon stimulation with epidermal growth factor (EGF). Photobleaching and photoactivation of Dronpa and fluorescence measurements were performed with laser-scanning confocal microscopy. ERK1-Dronpa was initially distributed throughout the cytosol and nucleus (Fig. 2A). After fluorescence was erased to background levels with a strong 488-nm laser line (1.4 W/cm^2), Dronpa was photoactivated in part of the cytoplasm (Fig. 2B). As a result of the high quantum yield of the photoactivation process (Φ_{PA} = 0.37), the marking required only a single 200-ms scan using a 405-nm laser line (0.20 W/cm^2). Next, a series of images was acquired by using a weak 488-nm laser line (0.014 W/cm^2) (Fig. 2D). Within 40 s, a substantial gradient of fluorescence signal was apparent across the nuclear envelope, which suggests inefficient nuclear import of ERK1-Dronpa. After erasure, nuclear export was examined by tracing the fluorescence of ERK1-Dronpa photoactivated in a region inside the nucleus (Fig. 2B). Again, inefficient transport was observed (Fig. 2F).

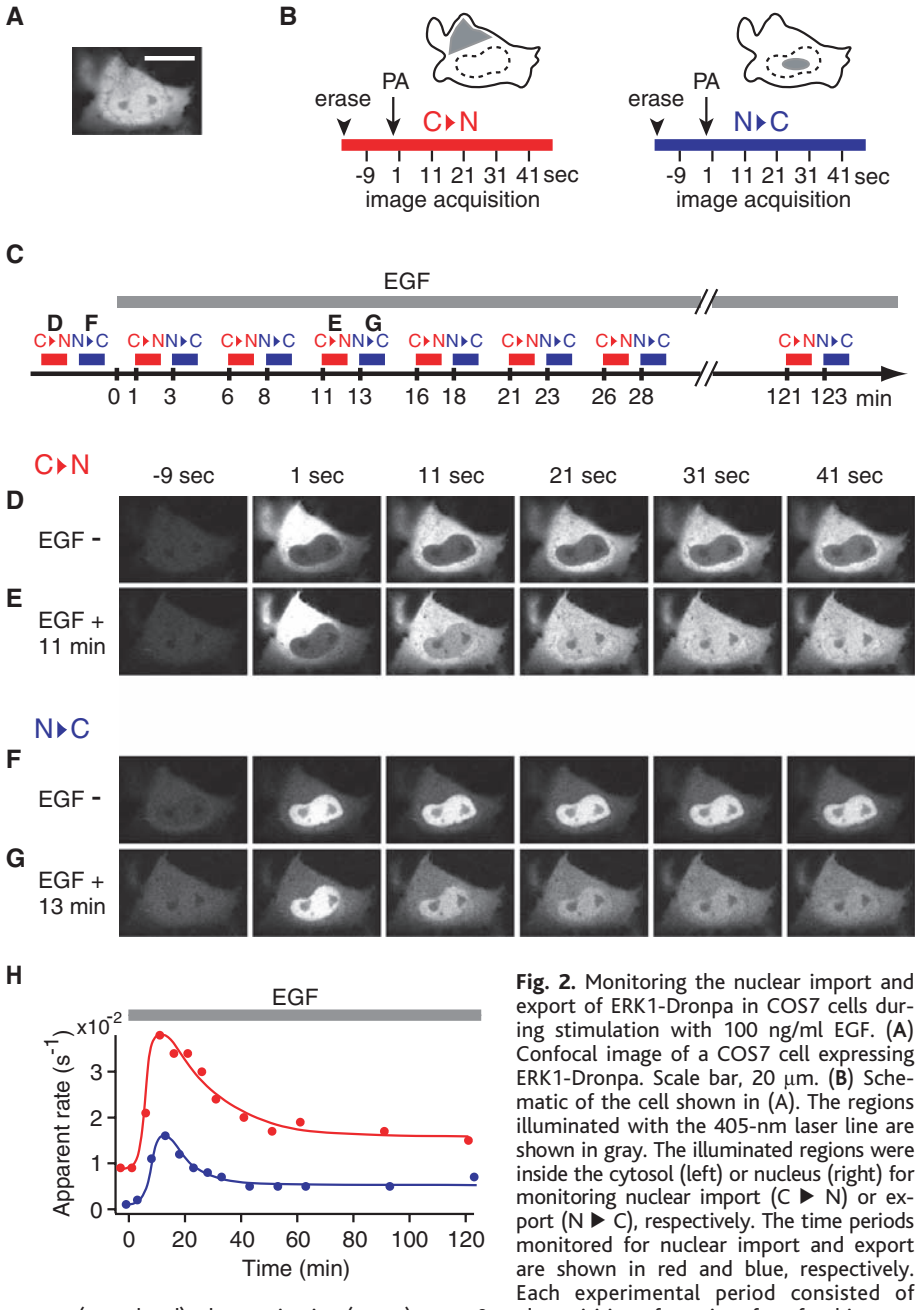


Fig. 2. Monitoring the nuclear import and export of ERK1-Dronpa in COS7 cells during stimulation with 100 ng/ml EGF. (**A**) Confocal image of a COS7 cell expressing ERK1-Dronpa. Scale bar, 20 μm. (**B**) Schematic of the cell shown in (A). The regions illuminated with the 405-nm laser line are shown in gray. The illuminated regions were inside the cytosol (left) or nucleus (right) for monitoring nuclear import (C ▶ N) or export (N ▶ C), respectively. The time periods monitored for nuclear import and export are shown in red and blue, respectively. Each experimental period consisted of erasure (arrowhead), photoactivation (arrow) at t = 0, and acquisition of a series of confocal images. (**C**) Timetable of intermittent monitoring of the nuclear import and export of ERK1-Dronpa before and during stimulation with EGF (gray bar). (**D** and **E**) Nuclear import of ERK1-Dronpa before stimulation (D) or after an 11-min incubation with EGF (E). (**F** and **G**) Nuclear export of ERK1-Dronpa before stimulation (F) or after a 13-min incubation with EGF (G). (**H**) Time courses of the nuclear influx (red) and efflux (blue) rates (*8*) of ERK1-Dronpa during EGF stimulation.

Next, these experiments were repeated eight times during continuous stimulation of the cell with 100 ng/ml EGF (Fig. 2C and movie S1). At early time points, the nucleocytoplasmic shuttling remained slow; however, 11 min after the onset of EGF stimulation, nuclear import was greatly enhanced (Fig. 2E). Interestingly, at a similar time point (13 min), nuclear export was also facilitated (Fig. 2G). This result contradicts a simple model in which a decrease in nuclear export accounts for nuclear accumulation of ERK. The rate of reduction of the fluorescence gradient across the nuclear envelope was quantified and plotted against time (Fig. 2H). The translocation of ERK1-Dronpa across the nuclear envelope was accelerated in both directions after several minutes of EGF stimulation. The phosphorylation of ERK1-Dronpa and its nuclear accumulation were confirmed by Western blotting and immunocytochemistry (fig. S3). The acceleration of the bidirectional flow of ERK1-Dronpa was also evident with lower concentrations of EGF (10 ng/ml), with the shuttling rate peaking with reduced amplitudes at similar or slightly later time points (fig. S4). Similar regulation was observed for ERK2-Dronpa in COS7 cells and HeLa cells that were stimulated with EGF. Thus, ERK signaling in the nucleus may be regulated by the rate of ERK nucleocytoplasmic shuttling. Assuming that ERK principally undergoes inactivation within the nucleus (*19*), fast circulation across the nuclear envelope is predicted to more effectively increase gene expression than does simple nuclear retention. Notably, variation in the initial shuttling rate between different cells was observed (fig. S5). Thus, any changes in movement must be assessed using data from a single cell, because measurements are affected by the geometry of the cells and marked regions.

Unlike ERK, nuclear localization signal (NLS)–containing proteins are transported by importins (*20*). The nuclear import of signal transduction and transcription factors, such as NF-κB and NF-AT, uses the NLS system. Importin β shuttles between the cytosol and nucleus as a translocation component by directly interacting with the nuclear pore complex (NPC). To observe the nucleocytoplasmic shuttling of importin β, we tagged importin β with Dronpa (importin-β-Dronpa). The chimeric protein was distributed throughout the cytosol and nucleus in COS7 cells, with some protein observed at the nuclear envelope (Fig. 3A). After erasure, Dronpa was photoactivated in a region of the cytosol (Fig. 3B) or nucleus (Fig. 3C) at $t = 0$, and a series of confocal images was acquired. Intense fluorescence at the nuclear envelope appeared immediately after photoactivation inside the cytosol or nucleus (Fig. 3, B and C), which suggests that both nuclear import and export of importin β involved direct interactions with the NPC. Because growth-factor stimulation changes the permeability of the nuclear envelope to NLS-containing gold particles (*21*), we quantified the shuttling rate of importin-β-Dronpa using the same protocol as for ERK1-Dronpa. The bidirectional flow rates were constant during incubation with 100 ng/ml EGF for 30 min (Fig. 3D), which suggests that the intrinsic nucleocytoplasmic shuttling of importin β was indifferent to growth-factor stimulation. The shuttling rate was variable between three different cells, stressing again the requirement for protein movement to be measured at multiple time points in individual cells.

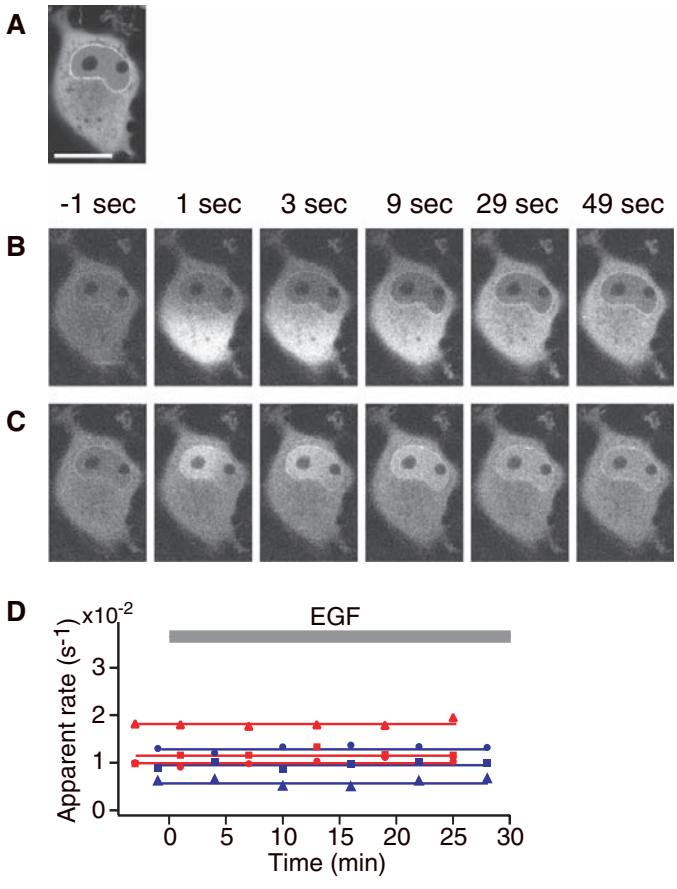


Fig. 3. Monitoring the nuclear import and export of importin-β-Dronpa in COS7 cells. (**A**) Confocal image of a COS7 cell expressing importin-β-Dronpa. Scale bar, 20 μm. (**B** and **C**) Nuclear import (B) and export (C) of importin-β-Dronpa. (**D**) Time courses of the nuclear influx (red) and efflux (blue) rates (*8*) of importin-β-Dronpa during stimulation with 100 ng/ml EGF, obtained from different three cells.

References and Notes

1. R. Y. Tsien, *Annu. Rev. Biochem.* **67**, 509 (1998).
2. J. Zhang, R. E. Campbell, A. Y. Ting, R. Y. Tsien, *Nature Rev. Mol. Cell Biol.* **3**, 906 (2002).
3. V. V. Verkhusha, K. A. Lukyanov, *Nature Biotechnol.* **22**, 289 (2004).
4. J. Lippincott-Schwartz, E. Snapp, A. Kenworthy, *Nature Rev. Mol. Cell Biol.* **2**, 444 (2001).
5. G. H. Patterson, J. Lippincott-Schwartz, *Science* **297**, 1873 (2002).
6. R. Ando, H. Hama, M. Yamamoto-Hino, H. Mizuno, A. Miyawaki, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12651 (2002).
7. D. M. Chudakov *et al.*, *Nature Biotechnol.* **21**, 191 (2003).
8. Materials and methods are available as supporting material on *Science* Online.
9. Y. A. Labas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4256 (2002).
10. R. E. Campbell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7877 (2002).
11. M. Irie, T. Fukaminato, T. Sasaki, N. Tamai, T. Kawai, *Nature* **420**, 759 (2002).
12. Y. C. Liang, A. S. Dvornikov, P. M. Rentzepis, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8109 (2003).
13. R. M. Dickson, A. B. Cubitt, R. Y. Tsien, W. E. Moerner, *Nature* **388**, 355 (1997).
14. A. Miyawaki, R. Y. Tsien, *Methods Enzymol.* **327**, 472 (2000).
15. T. S. Lewis, P. S. Shapiro, N. G. Ahn, *Adv. Cancer Res.* **74**, 49 (1998).
16. M. J. Robinson, M. H. Cobb, *Curr. Opin. Cell Biol.* **9**, 180 (1997).
17. T. Furuno, N. Hirashima, S. Onizawa, N. Sagiya, M. Nakanishi, *J. Immunol.* **166**, 4416 (2001).
18. A. M. Horgan, P. J. S. Stork, *Exp. Cell Res.* **285**, 208 (2003).
19. V. Volmat, M. Camps, S. Arkinstall, J. Puoyssegur, P. Lenormand, *J. Cell Sci.* **114**, 3433 (2001).
20. L. Xu, J. Massagué, *Nature Rev. Mol. Cell Biol.* **5**, 209 (2004).
21. C. M. Feldherr, D. Akin, *Exp. Cell Res.* **205**, 179 (1993).

22. We thank S. Karasawa for preparation of coral and for valuable advice, M. Cobb for the ERK1 construct, Y. Yoneda for the human importin β1 construct, and S. Habuchi, J. Hofkens, Y. Gotoh, N. Imamoto, Y. Sato, K. Takishima, and C. Bargmann for critical comments. This work was partly supported by grants from the Japanese Ministry of Education, Science, and Technology, from the Human Frontier Science Program, and from the New Energy and Industrial Technological Development Organization.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1370/DC1
Materials and Methods
Figs. S1 to S5
Table S1
Movie S1

8 July 2004; accepted 3 September 2004

TRPM4 Regulates Calcium Oscillations After T Cell Activation

Pierre Launay,[1]*† Henrique Cheng,[2]*‡ Subhashini Srivatsan,[1] Reinhold Penner,[2] Andrea Fleig,[2] Jean-Pierre Kinet[1]§

TRPM4 has recently been described as a calcium-activated nonselective (CAN) cation channel that mediates membrane depolarization. However, the functional importance of TRPM4 in the context of calcium (Ca^{2+}) signaling and its effect on cellular responses are not known. Here, the molecular inhibition of endogenous TRPM4 in T cells was shown to suppress TRPM4 currents, with a profound influence on receptor-mediated Ca^{2+} mobilization. Agonist-mediated oscillations in intracellular Ca^{2+} concentration ($[Ca^{2+}]_i$), which are driven by store-operated Ca^{2+} influx, were transformed into a sustained elevation in $[Ca^{2+}]_i$. This increase in Ca^{2+} influx enhanced interleukin-2 production. Thus, TRPM4-mediated depolarization modulates Ca^{2+} oscillations, with downstream effects on cytokine production in T lymphocytes.

The intensity and shape of Ca^{2+} signals (e.g., oscillations versus sustained Ca^{2+} influx) have been recognized to be essential in setting the threshold for different transcription programs (*1*). This has been particularly well established in the context of T lymphocyte activation and NFAT (nuclear factor of activated T cells)–dependent interleukin-2 (IL-2) production (*2*–*4*). The molecular and electrophysiological characterization of TRPM4 as a Ca^{2+}-activated nonselective (CAN) channel (*5*–*8*) raised the question of its potential role in modulating Ca^{2+}-dependent cellular responses in the physiological context of intact cells. Two splice variants of TRPM4 are known: the short form, TRPM4a (*9*), which lacks 174 amino acid residues at the N terminus; and the long form, TRPM4b (*5*) (referred to here as TRPM4). The latter is a widely expressed CAN channel (*10*, *11*) that does not conduct Ca^{2+} but instead mediates cell membrane depolarization. We hypothesized that the capacity of TRPM4 to depolarize the cells in response to changes in $[Ca^{2+}]_i$ can affect the shape and intensity of Ca^{2+} signals initiated by receptor stimulation, thereby modulating important effector functions such as the transcription programs leading to IL-2 production.

Multimerization of TRPM4 is likely required to form the proper pore structure of the channel. Therefore, we designed a strategy to suppress TRPM4 currents using a dominant negative approach. Inactive monomers of mutants with an intact capacity to multimerize could associate with endogenous channels and generate a dominant negative effect. This would permit the analysis of the role of TRPM4 in effector functions. Two deletion mutants of TRPM4 were generated and analyzed (*12*): deletion of the first 177 amino acids in the N terminus (ΔN-TRPM4) and of the last 160 amino acids in the C terminus (ΔC-TRPM4). The C terminus contains a coil-coiled domain that could be involved in protein-protein interactions (Fig. 1A). The cDNAs corresponding to the Flag-tagged wild-type (wt)-TRPM4 and the two V5-tagged deleted forms were each transfected into a human embryonic kidney 293 (HEK-293) cell line.

Selected stable clones overexpressing either wt-, ΔN-, or ΔC-TRPM4 were assessed for their capacity to localize to the cell surface. The cells were subjected to surface iodination, and the lysates were immunoprecipitated with anti-V5, anti-Flag, and

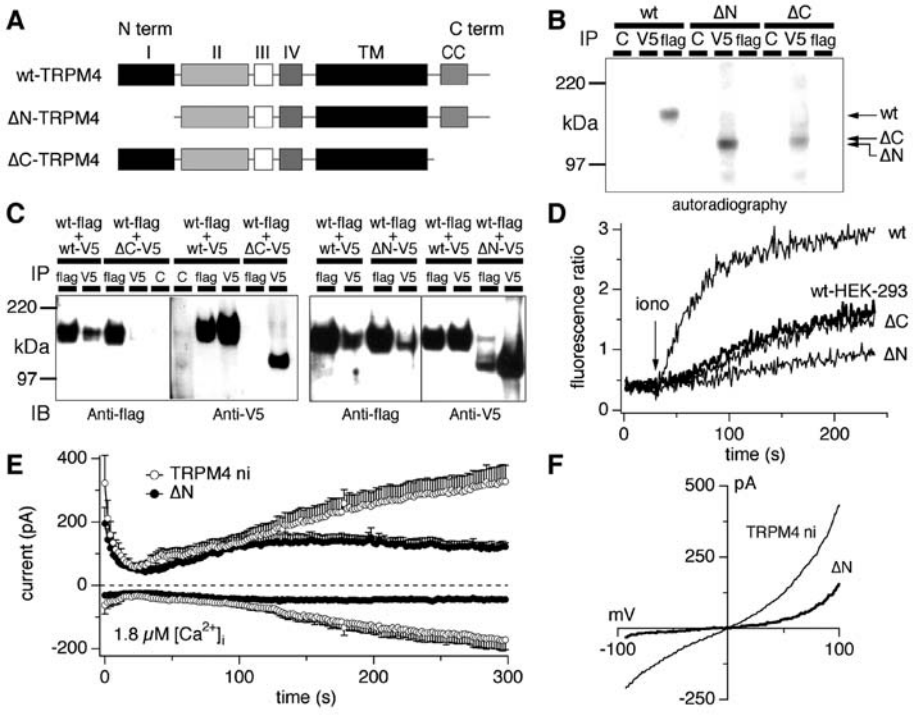


Fig. 1. Structure-function analysis of TRPM4 in HEK-293 cells. (**A**) Schematic representation of wt-TRPM4 and the truncated forms ΔN-TRPM4 and ΔC-TRPM4 with the N-terminal unique regions I to IV, transmembrane domain (TM), and coiled-coil region (CC). (**B**) Surface expression of wt-TRPM4 and the truncated forms. Surface-expressed proteins are labeled with iodine. The wt-TRPM4 molecule was immunoprecipitated with anti-Flag and the two truncated forms, ΔC and ΔN, with anti-V5. (**C**) The C-terminal domain of TRPM4 is involved in its multimerization. HEK-293 cells were cotransfected with two different tagged forms (V5 and Flag) of wt-TRPM4 or cotransfected with Flag-tagged wt-TRPM4 and V5-tagged ΔN-TRPM4 or V5-tagged ΔC-TRPM4. Proteins of cell lysates were immunoprecipitated with Flag and V5, and filters of immune complexes were blotted with both anti-V5 and anti-Flag. (**D**) Na^+ entry following TRPM4 activation by ionomycin (2.5 μM) in intact HEK-293 cell populations loaded with SBFI. Both wild-type HEK-293 cells, as well as cells transfected with ΔC-TRPM4, produced intermediate Na^+ entry, reflecting activity of endogenous TRPM4. Tetracycline-induced cells overexpressing wt-TRPM4 produced larger Na^+ entry, whereas cells expressing the truncated ΔN-TRPM4 form had reduced Na^+ entry. (**E**) Average inward and outward currents at –80 and +80 mV in noninduced (TRPM4 ni) and ΔN-TRPM4–expressing HEK-293 cells stimulated by $[Ca^{2+}]_i$ clamped at 1.8 μM (n = 6 to 7 cells, mean ± SEM). (**F**) *I-V* relationships under the same experimental conditions as in (E), obtained from representative cells 300 s after whole-cell establishment.

isotype control (C) antibodies. Specific antibodies (anti-Flag or anti-V5) precipitated single bands with the predicted molecular sizes (134 kD for wt-TRPM4, 115 kD for ΔN-TRPM4, and 118 kD for ΔC-TRPM4), and no band was seen with the nonspecific or isotype control antibodies (Fig. 1B). These data indicate that both ΔN- and ΔC-TRPM4 are expressed at the plasma membrane.

The ability of ΔN- or ΔC-TRPM4 to heteromultimerize with wt-TRPM4 was assessed by expressing the relevant constructs in HEK-293 cells. Flag-tagged wt-TRPM4 was expressed with either V5-tagged wt-, ΔN-, or ΔC-TRPM4. Flag–wt-TRPM4 was immunoprecipitated in complexes with V5-wt-TRPM4 and with V5-ΔN-TRPM4 but not with V5-ΔC-TRPM4 (Fig. 1C), indicating that monomers of wt-TRPM4 can homomultimerize and that ΔN-TRPM4, but not ΔC-TRPM4, can heteromultimerize with wt-TRPM4. Thus, it appears that the last 160 amino acids of the C-terminal region of TRPM4 may be required for multimerization.

Because the main ion carried by TRPM4 is Na^+, we tested the capacity of the mutant and wt-TRPM4 channels to mediate Na^+ entry in response to an ionomycin-induced elevation in $[Ca^{2+}]_i$ (*12*), using sodium-binding benzofuran isophthalate (SBFI), a specific ratiometric Na^+ dye (*13*, *14*). Ionomycin application induced an increase in $[Na^+]_i$ in wt-HEK-293 cells, presumably through endogenous TRPM4. An even greater increase was observed in cells overexpressing wt-TRPM4 (wt) (Fig. 1D). However, expression of ΔC-TRPM4 did not affect the endogenous cation channels, suggesting that the C terminus is involved in TRPM4 multimerization. In contrast, there was reduced Na^+ entry in cells expressing ΔN-TRPM4, suggesting a dominant negative effect of ΔN-TRPM4 on endogenous TRPM4.

To determine whether the reduction in $[Na^+]_i$ observed in the ΔN-TRPM4–expressing cells was a direct consequence of the inhibition of endogenous TRPM4, we used patch-clamp recordings (*12*) to assess whole-cell membrane currents in noninduced HEK-293 cells and in ΔN-TRPM4 transfectants with $[Ca^{2+}]_i$ clamped at 1.8 μM (Fig. 1E). The noninduced cells generated currents with current-voltage (*I-V*) characteristics similar to those of endogenous TRPM4 described previously (*5*) (Fig. 1F). In contrast, TRPM4 currents in ΔN-TRPM4 transfectants were greatly diminished (Fig. 1, E and F), consistent with the interpretation that ΔN-TRPM4 has a dominant negative effect on endogenous TRPM4 currents.

To determine the function of TRPM4, we examined endogenous channels in immune cells, the cellular responses of which heavily rely on long-lasting oscillatory changes in $[Ca^{2+}]_i$. To assess the endogenous expression of the TRPM4 molecule, we generated a rabbit polyclonal antibody against TRPM4. The antibody recognized V5-wt-TRPM4 by immunoprecipitation and Western blot analysis (*15*). Anti-TRPM4 also detected endogenous TRPM4 protein in HEK-293 cells, in which we originally characterized TRPM4 (*5*) (Fig. 2A). It also recognized TRPM4 in mouse thymocytes, in the mouse Th2 cell clone, D10.G4, and in the two human T lymphoblast cell lines, Molt-4 and Jurkat (Fig. 2B), suggesting that TRPM4 is a conserved and an important ion channel for lymphocytes across multiple species. The antibody reacts specifically with TRPM4, because the single band detected by the antibody disappeared when Jurkat cell lysates were preincubated with the immunizing peptide (Fig. 2C).

Jurkat T cells are a widely used model for the study of Ca^{2+} signaling (*4*, *16*, *17*). Perfusion with 100 to 800 nM $[Ca^{2+}]_i$ revealed concentration-dependent activation of large currents (Fig. 2D). The *I-V* relationships obtained at 200 s after initiation of the experiment for each concentration (Fig. 2E) are consistent with those previously reported for TRPM4. Because TRPM4 is a monovalent-specific cation channel that does not pass any appreciable amount of Ca^{2+}, we assessed whether Na^+ is the main cation responsible for inward currents. When extracellular Na^+ was substituted with isotonic *N*-methyl-D-glucamine chloride (NMDG), almost complete abolition of the inward currents was observed, whereas the outward currents were only slightly reduced. This effect was reversible when NaCl-based Ringer solution was readmitted (Fig. 2F). The *I-V* relationships of

[1]Department of Pathology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02215, USA. [2]Laboratory of Cell and Molecular Signaling, Center for Biomedical Research at The Queen's Medical Center and John A. Burns School of Medicine at the University of Hawaii, Honolulu, HI 96813, USA.

*These authors contributed equally to this work.
†Present address: INSERM E0225, Bichat Medical School, 75870 Paris Cedex 18, France.
‡Present address: All Children's Hospital and University of South Florida, 140 Seventh Avenue South, CRI 2012 St. Petersburg, FL 33701, USA.
§To whom correspondence should be addressed. E-mail: jkinet@bidmc.harvard.edu

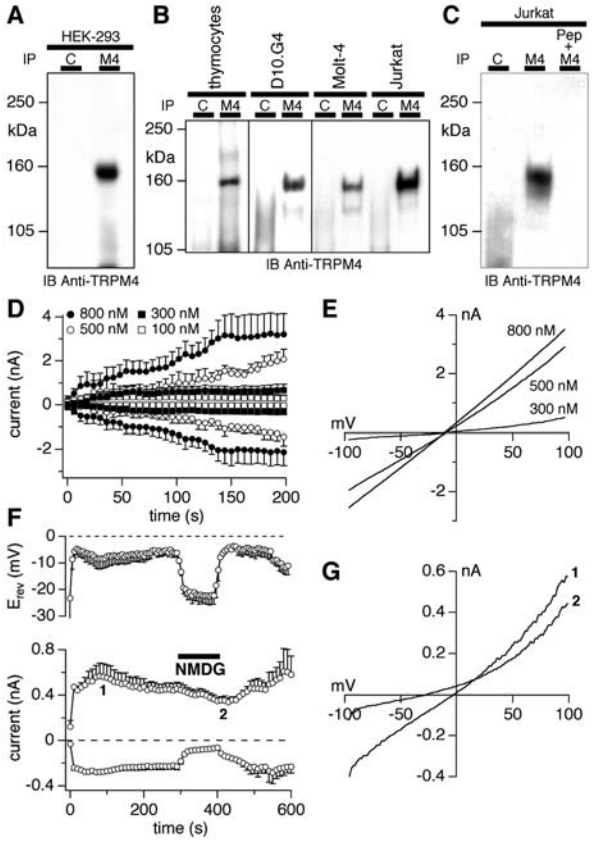


Fig. 2. Characterization of TRPM4 in Jurkat T lymphocytes. (**A**) HEK-293 cells were analyzed for the endogenous expression of TRPM4 protein after immunoprecipitation and immunoblotting with the polyclonal antibody against TRPM4 (M4). Control immunoprecipitation with an irrelevant antibody is indicated by C. (**B**) Detection of endogenous TRPM4 after immunoprecipitation and immunoblotting with anti-TRPM4 in mouse thymocytes and the mouse Th2 clone D10.G4, as well as in the human T cell lines Molt-4 and Jurkat. (**C**) The anti-TRPM4 reactivity is inhibited by preincubation with the peptide. (**D**) Endogenous TRPM4-like currents in Jurkat cells. Average inward and outward currents in wt-Jurkat cells at –80 and +80 mV with $[Ca^{2+}]_i$ clamped at the indicated concentrations (n = 3 to 5 cells, mean ± SEM). (**E**) *I-V* relationship under the same experimental conditions as in (D), obtained from representative cells 200 s after whole-cell establishment. (**F**) (Bottom panel) Average inward and outward currents carried by TRPM4 at –80 and +80 mV, respectively. Cells were perfused with solutions in which $[Ca^{2+}]_i$ was buffered at 800 nM (n = 3 cells, mean ± SEM). For the time indicated, cells were exposed to an isotonic NMDG-Cl–based solution that additionally contained 1 mM $CaCl_2$ and 2 mM $MgCl_2$. The numbers 1 and 2 indicate the time at which raw data traces displayed in (F) were extracted. (Top panel) Reversal potentials extracted from individual ramp current records. (**G**) *I-V* relationships of TRPM4 currents under the same experimental conditions as in (F), measured in a representative cell before (1) and during (2) application of an NMDG-Cl–based solution.

whole-cell currents obtained at the peak of TRPM4 activation and during exposure to NMDG confirm that the reduction in Na^+ influx causes a hyperpolarizing shift in reversal potential (Fig. 2F, top panel, and 2G). Taken together, these data demonstrate that Jurkat cells express functional endogenous CAN channels with the same characteristics as those described previously for TRPM4.

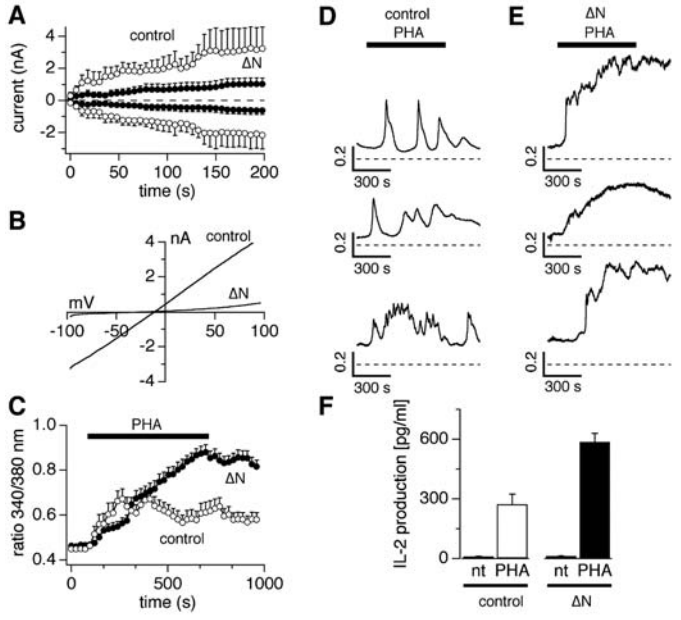


Fig. 3. Structural dissociation of TRPM4 increases receptor-mediated Ca^{2+} entry and IL-2 production. (**A**) Average inward and outward currents in wt-Jurkat and ΔN-TRPM4–expressing cells at −80 and +80 mV with $[Ca^{2+}]_i$ clamped at 800 nM (n = 5 to 6 cells, mean ± SEM). The dominant negative effect of ΔN-TRPM4 is apparent. (**B**) *I-V* relationships under the same experimental conditions as in (A), obtained from representative cells 200 s after whole-cell establishment. (**C**) Average $[Ca^{2+}]_i$ signals from ΔN-TRPM4 (n = 45 cells, 4 experiments) and wt-Jurkat cells (n = 25 cells, 3 experiments) loaded with Fura-2-AM and stimulated with PHA (20 μg/ml). (**D** and **E**) Representative traces of three individual wt-Jurkat and ΔN-TRPM4 Jurkat cells. (**F**) PHA-induced IL-2 production in control wt-Jurkat and ΔN-TRPM4-Jurkat cells. The data shown are the means (±SD) of three independent experiments ($P < 0.004$, unpaired Student's t test).

Clones of Jurkat cells expressing ΔN-TRPM4 were selected, based on protein expression (*15*), and compared to clones expressing the empty vector (control). Cells were perfused with 800 nM $[Ca^{2+}]_i$. Inward currents, obtained at –80 mV after 200 s, in the ΔN-TRPM4 clones were on average ~600 pA, which is smaller than those seen in control cells by a factor of about 3 (Fig. 3, A and B). Thus, the expression of ΔN-TRPM4 has a strong dominant negative effect on endogenous TRPM4 currents in lymphocytes. Conversely, expression of ΔC-TRPM4 in Jurkat cells had no effect on endogenous TRPM4 currents, and these currents were comparable to the ones from clones expressing the empty vector (*15*).

Given that TRPM4 activation would lead to membrane depolarization, we studied the effect of TRPM4 on Ca^{2+} influx in both wt-Jurkat cells and cells in which TRPM4 was suppressed by the ΔN-TRPM4 construct. Average Ca^{2+} signals from ΔN-TRPM4 (n = 45 cells, 4 experiments) and control wt-Jurkat cells (n = 25 cells, 3 experiments) stimulated with phytohemagglutinin (PHA, 20 μg/ml) were determined (Fig. 3C). Representative traces from wt-control cells revealed a pattern of changes in $[Ca^{2+}]_i$ that was characterized by oscillations (Fig. 3D). In ΔN-TRPM4 cells, the oscillatory pattern in $[Ca^{2+}]_i$ was transformed into a prolonged sustained Ca^{2+} influx (Fig. 3E). IL-2 cytokine release by lymphocytes (*12*) is dependent on increases in $[Ca^{2+}]_i$, and IL-2 secretion of PHA-stimulated ΔN-TRPM4 cells was increased 2.2-fold compared with that of the control wt-Jurkat cells (Fig. 3F).

To further evaluate TRPM4 regulation of Ca^{2+} influx in Jurkat T cells and to eliminate possible clonal variations resulting from cellular drug selection, we reduced expression of endogenous TRPM4 using an RNA interference (RNAi) approach (*12*, *18*). About 90 to 95% of Jurkat cells were positively infected as determined with a green fluorescent protein (GFP)–control retroviral vector (*15*). TRPM4-specific siRNA decreased TRPM4 mRNA (Fig. 4A, upper panel) as compared to treatment with a scrambled sequence. There was no decrease in the level of small ribosomal protein (Fig. 4A, lower panel). Immunoprecipitation and Western blot analysis with the polyclonal antibody to TRPM4 (Fig. 4B, upper panel) showed a decrease in TRPM4 protein in the TRPM4 siRNA–infected Jurkat cells. No changes in β-actin protein levels were observed (Fig. 4B, lower panel). Electrophysiology analysis further verified the decrease of TRPM4 currents in TRPM4 siRNA–infected Jurkat cells (Fig. 4, C and D).

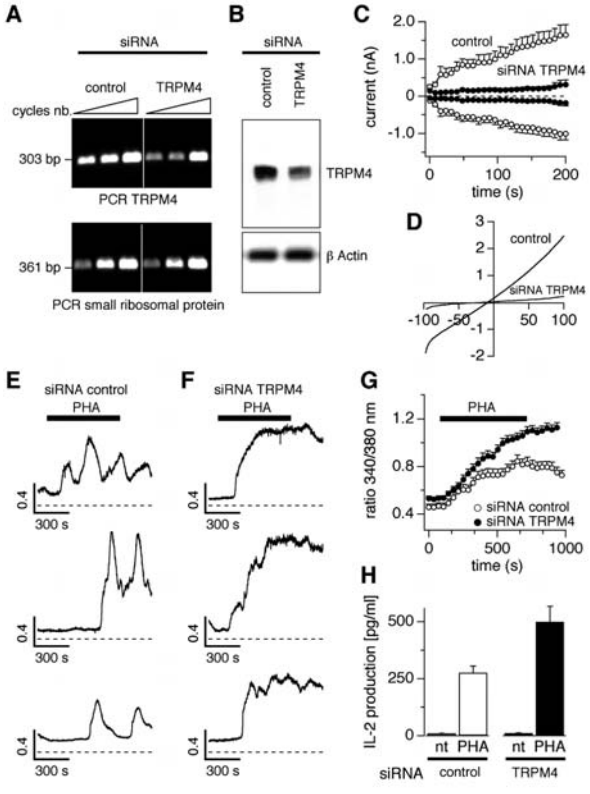


Fig. 4. Gene silencing of TRPM4 increases receptor-mediated Ca^{2+} entry and IL-2 production. (**A**) Reverse transcription–polymerase chain reaction (RT-PCR) of TRPM4 mRNA from Jurkat cells infected with a TRPM4-specific siRNA and a scrambled sequence control. Number of cycles: 32, 34, and 36. Positive control for RT-PCR used primers specific for small ribosomal protein. Number of cycles: 22, 24, and 26. (**B**) Cell lysates were immunoprecipitated with anti-TRPM4 and anti–β-actin and subjected to Western blotting with the same antibodies, respectively. (**C**) Average inward and outward currents in Jurkat cells infected with a TRPM4-specific siRNA (n = 6 ± SEM) and a scrambled sequence control (n = 7 ± SEM) at −80 and +80 mV with $[Ca^{2+}]_i$ clamped at 800 nM. The decrease in current due to the TRPM4 siRNA effect is apparent. (**D**) *I-V* relationships obtained from representative cells 200 s after whole-cell establishment. (**E** and **F**) Representative $[Ca^{2+}]_i$ signals from control siRNA and TRPM4 siRNA Jurkat cells. (**G**) Average Ca^{2+} signals from TRPM4 siRNA (n = 64 cells, 5 experiments) and control siRNA (n = 33 cells, 4 experiments) cells loaded with Fura-2-AM and stimulated with PHA (20 μg/ml). (**H**) PHA-induced IL-2 production in control siRNA and TRPM4 siRNA–infected Jurkat cells. The data shown are the means (±SD) of five independent experiments ($P < 0.017$, unpaired Student's t test).

The siRNA-infected cells were assessed for changes in Ca^{2+} oscillation patterns. Average Ca^{2+} signals from Jurkat cells infected with TRPM4 siRNA (n = 64 cells, 5 experiments) and control siRNA (n = 33 cells, 4 experiments) were determined (Fig. 4G). Jurkat cells infected with TRPM4 siRNA exhibited a more prolonged sustained Ca^{2+} influx as compared to cells infected with the siRNA control, which exhibit oscillatory changes typical of wt-Jurkat cells (compare Fig. 4, E and F). PHA stimulation of these cells resulted in a twofold increase in the amount of IL-2 secreted by the TRPM4 siRNA–infected cells as compared with the amount of IL-2 secreted by the control cells (Fig. 4H).

In summary, our results establish that TRPM4 is a previously unrecognized ion channel in Jurkat T cells with a profound influence on Ca^{2+} signaling. Molecular suppression of TRPM4 converts oscillatory changes of $[Ca^{2+}]_i$ into long-lasting sustained elevations in Ca^{2+} and leads to augmented IL-2 production. It is conceivable that this effect occurs physiologically in cells that express the short splice variant TRPM4a, which is of a similar length to ΔN-TRPM4 and could therefore act as a native dominant negative subunit. In electrically nonexcitable cells, TRPM4 would tend to reduce Ca^{2+} influx by depolarizing the membrane potential and reducing the driving force for Ca^{2+} entry through store-operated CRAC (Ca^{2+} release–activated Ca^{2+}) channels.

The molecular and electrophysiological identification of TRPM4 in Jurkat T cells may call for a reinterpretation of the interplay of ionic currents that shape intracellular Ca^{2+} signals (*4*, *16*, *19*). We propose that TRPM4 acts in concert with CRAC, Kv1.3, and K_{Ca} channels to control $[Ca^{2+}]_i$ oscillations in lymphocytes through oscillatory changes in membrane potential according to the following model.

At rest, the lymphocyte membrane potential is around –60 mV, owing to the basal activity of K^+ channels (*20*). Engagement of T cell receptors induces phospholipase C–mediated production of $InsP_3$ (inositol 1,4, 5-trisphosphate), which causes Ca^{2+} release and activation of store-operated CRAC channels. Current models of Ca^{2+} oscillations in lymphocytes (*19*) propose that the I_{CRAC}-mediated Ca^{2+} influx triggers the activation of Ca^{2+}-activated K^+ channels, which provides the driving force for Ca^{2+} entry by hyperpolarizing the membrane potential until $[Ca^{2+}]_i$ reaches a high-enough level to inhibit I_{CRAC}. As Ca^{2+} entry through I_{CRAC} is reduced, $[Ca^{2+}]_i$ falls until it reaches a level that removes the negative feedback on I_{CRAC}, and the cycle resumes by increasing Ca^{2+} entry through I_{CRAC}. This model lacks a strong depolarizing conductance that would be required to recruit voltage-dependent Kv1.3 channels present in T cells and could account for the observed oscillations in membrane potential (*20*, *21*). The $[Ca^{2+}]_i$-dependent activation of TRPM4 channels may provide this mechanism by becoming activated at around the peak of an oscillatory Ca^{2+} transient, causing the membrane potential to depolarize and thereby substantially reducing the driving force for Ca^{2+} influx. The depolarization would then recruit voltage-dependent K^+ currents (Kv1.3), which would tend to repolarize the membrane potential and also aid in the closure of TRPM4 channels, because the open probability of TRPM4 channels is reduced at negative membrane voltages (*5*, *6*, *8*). The repolarization would reestablish the driving force for Ca^{2+} influx through I_{CRAC} so that the next oscillation in $[Ca^{2+}]_i$ can take place.

References and Notes

1. G. R. Crabtree, *J. Biol. Chem.* **276**, 2313 (2001).
2. R. E. Dolmetsch, K. Xu, R. S. Lewis, *Nature* **392**, 933 (1998).
3. W. Li, J. Llopis, M. Whitney, G. Zlokarnik, R. Y. Tsien, *Nature* **392**, 936 (1998).
4. R. S. Lewis, *Annu. Rev. Immunol.* **19**, 497 (2001).
5. P. Launay *et al.*, *Cell* **109**, 397 (2002).
6. T. Hofmann, V. Chubanov, T. Gudermann, C. Montell, *Curr. Biol.* **13**, 1153 (2003).
7. M. Murakami *et al.*, *Biochem. Biophys. Res. Commun.* **307**, 522 (2003).
8. B. Nilius *et al.*, *J. Biol. Chem.* **278**, 30813 (2003).
9. X. Z. Xu, F. Moebius, D. L. Gill, C. Montell, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10692 (2001).
10. O. H. Petersen, *Curr. Biol.* **12**, R520 (2002).
11. B. Nilius, G. Droogmans, R. Wondergem, *Endothelium* **10**, 5 (2003).
12. Materials and methods are available as supporting material on *Science* Online.
13. A. T. Harootunian, J. P. Kao, B. K. Eckert, R. Y. Tsien, *J. Biol. Chem.* **264**, 19458 (1989).
14. A. Minta, R. Y. Tsien, *J. Biol. Chem.* **264**, 19449 (1989).
15. P. Launay *et al.*, data not shown.
16. M. D. Cahalan, H. Wulff, K. G. Chandy, *J. Clin. Immunol.* **21**, 235 (2001).
17. R. T. Abraham, A. Weiss, *Nat. Rev. Immunol.* **4**, 301 (2004).
18. T. R. Brummelkamp, R. Bernards, R. Agami, *Science* **296**, 550 (2002).
19. R. E. Dolmetsch, R. S. Lewis, *J. Gen. Physiol.* **103**, 365 (1994).
20. J. A. Verheugen, H. P. Vijverberg, *Cell Calcium* **17**, 287 (1995).
21. C. M. Fanger *et al.*, *J. Biol. Chem.* **276**, 12249 (2001).

22. We thank M. K. Monteilh-Zoller and C. E. Oki for technical assistance, L. Glimcher's laboratory for providing the murine T cell clone D10.G4, and S. Kraft for insightful advice on the RNAi method. This work was supported in part by NIH grants R01-AI46734 (J.-P.K.); R01-NS40927, R01-AI50200, and R01-GM63954 (R.P.); and R01-GM65360 (A.F.). P.L. was supported by a fellowship from the Human Frontier Science Program Organization.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1374/DC1
Materials and Methods

6 April 2004; accepted 14 September 2004

A Protein Sensor for siRNA Asymmetry

Yukihide Tomari, Christian Matranga, Benjamin Haley, Natalia Martinez, Phillip D. Zamore*

To act as guides in the RNA interference (RNAi) pathway, small interfering RNAs (siRNAs) must be unwound into their component strands, then assembled with proteins to form the RNA-induced silencing complex (RISC), which catalyzes target messenger RNA cleavage. Thermodynamic differences in the base-pairing stabilities of the 5′ ends of the two ~21-nucleotide siRNA strands determine which siRNA strand is assembled into the RISC. We show that in *Drosophila*, the orientation of the Dicer-2/R2D2 protein heterodimer on the siRNA duplex determines which siRNA strand associates with the core RISC protein Argonaute 2. R2D2 binds the siRNA end with the greatest double-stranded character, thereby orienting the heterodimer on the siRNA duplex. Strong R2D2 binding requires a 5′-phosphate on the siRNA strand that is excluded from the RISC. Thus, R2D2 is both a protein sensor for siRNA thermodynamic asymmetry and a licensing factor for entry of authentic siRNAs into the RNAi pathway.

In *Drosophila* lysates, siRNAs are loaded into the RISC by an ordered pathway in which one of the two siRNA strands, the guide strand, is assembled into the RISC, whereas the other strand, the passenger strand, is excluded and destroyed (*1*–*14*). A central step in RISC assembly is formation of the RISC-loading complex [RLC, previously designated complex A (*13*)], which contains double-stranded siRNA, the double-stranded RNA binding protein R2D2, and Dicer-2 (Dcr-2), as well as additional unidentified proteins. The function of Dicer in loading siRNA into the RISC is distinct from its role in generating siRNA from long double-stranded RNA (dsRNA) (*10*, *15*). Both R2D2 and Dcr-2 are

Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA.

*To whom correspondence should be addressed. E-mail: phillip.zamore@umassmed.edu

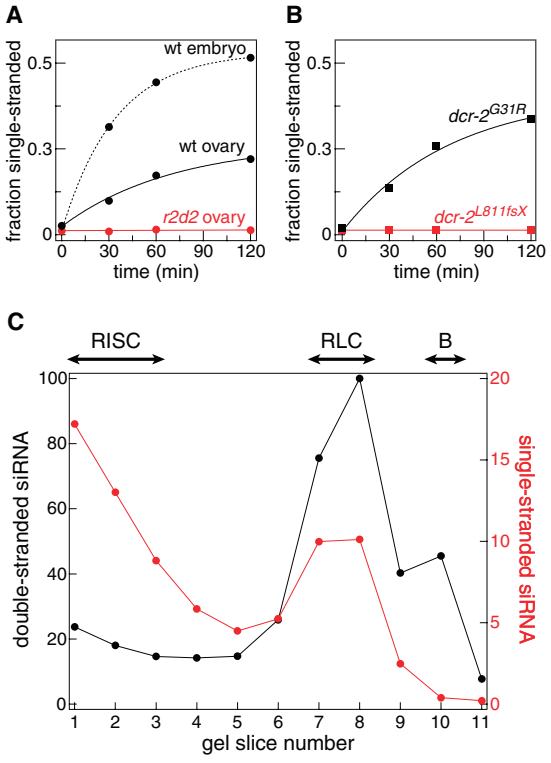


Fig. 1. The RISC-loading complex (RLC) initiates siRNA unwinding. (**A**) *r2d2* mutant ovary lysates, which cannot assemble RLC, do not unwind siRNA (wt, wild type). (**B**) siRNA unwinding was defective in ovary lysate from a mutation that disrupts all known Dcr-2 functions (L811fsX), including RLC assembly (*10, 12, 18*), but was normal for a point mutation in *dcr-2* ($Gly^{31} \rightarrow Arg$, G31R) that disrupts its function in dicing long dsRNA into siRNA, but not RISC assembly (*10, 12*). (**C**) The RLC, which is composed largely of double-stranded siRNA, also contains single-stranded siRNA. Note the different scales for the relative amounts of single- and double-stranded siRNA.

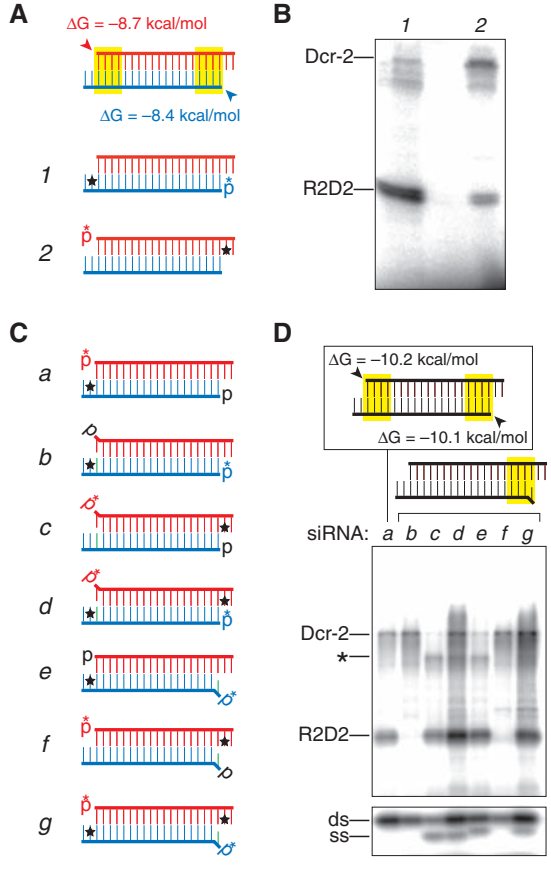


Fig. 2. Asymmetric binding of the Dcr-2/R2D2 heterodimer to siRNA duplexes in *Drosophila* embryo lysate. (**A**) Structure of the fully base-paired but asymmetric luciferase siRNA duplex used in (B) for protein-siRNA photocrosslinking. The local thermodynamic stability of the yellow highlighted base pairs is indicated here and in (D). Stars denote 5-iodouracil; an asterisk denotes a 5′-[^{32}P]phosphate. (**B**) The blue strand generates about 5 times as much RISC as the red; more Dcr-2 photocrosslinks to the 5-iodouracil nearest the less thermodynamically stable end, whereas more R2D2 photocrosslinks to the 5-iodouracil nearest the more thermodynamically stable end. (**C**) The series of 5-iodouracil substituted siRNAs used in (D). siRNA *a* is thermodynamically symmetric. siRNAs *b* to *g* contain a single unpaired nucleotide at the 5′ end of one strand, making them highly asymmetric (*7*). For siRNAs *b* to *d*, the red strand serves as the guide strand, whereas the blue strand serves as the guide for siRNAs *e* to *g*. (**D**) For the functionally asymmetric siRNAs, Dcr-2 binding was detected nearest the 5′ end of the guide strand, whereas R2D2 was detected near the 5′ end of the passenger strand, the more thermodynamically stable end. The lower panel shows the ^{32}P-radiolabeled siRNA strands at the bottom of the same gel. Single-stranded siRNA was detected only when the labeled strand served as a guide strand and entered the RISC. The asterisk indicates a photocrosslink to Ago2.

required to form RLC (*13*) and to unwind siRNA (Fig. 1, A and B), but recombinant Dcr-2/R2D2 heterodimer or Dcr-2 alone cannot catalyze siRNA unwinding (fig. S1). Thus, the Dcr-2/R2D2 heterodimer is necessary but not sufficient to unwind siRNA.

If siRNA unwinding is initiated in the RLC, then the RLC should contain some single-stranded siRNA. To test this idea, we briefly incubated siRNA duplex in lysate, resolved the complexes formed by native gel electrophoresis, divided the gel into 11 parts, and analyzed the structure of the siRNA in each gel slice (fig. S2A). Consistent with our previous findings, a peak of double-stranded siRNA comigrated with both the RLC and complex B, which is thought to be a precursor to RLC (*13*) (Fig. 1C). A small peak of single-stranded siRNA also comigrated with the RLC, but not with complex B (Fig. 1C), which suggests that the RLC initiates siRNA unwinding. Similar peaks of single-stranded siRNA comigrated with the RLC for the passenger strand of this siRNA and for the guide and passenger strands of a second siRNA (fig. S2B). We conclude that the RLC initiates siRNA unwinding.

The RLC also senses siRNA thermodynamic asymmetry, thereby determining which strand enters the RISC. siRNA containing 5-iodouracil at the 20th nucleotide (p20) can be photocrosslinked to R2D2 and Dcr-2 (*13*). Photocrosslinking is position-specific: An siRNA containing 5-iodouracil at position 12 was not cross-linked to R2D2 or Dcr-2 (*13*). Photocrosslinking attaches the radiolabel of the siRNA to the protein, identifying proteins that lie near p20 of the substituted siRNA strand. We evaluated the relative efficiency of photocrosslinking to R2D2 and Dcr-2 for three types of siRNA (Fig. 2) (table S1): a luciferase-specific siRNA whose sequence makes the 5′ end of the antisense strand less thermodynamically stable than the 5′ end of the sense strand; a nearly symmetric siRNA targeting human *Zn*, *Cu superoxide dismutase 1* (*sod1*), in which the stabilities of the 5′ ends are essentially the same; and a series of highly asymmetric *sod1*-directed siRNAs in which the first nucleotide of the guide strand is mismatched to the passenger strand, causing the guide strand to be loaded into the RISC almost exclusively. When we used the partially asymmetric luciferase-specific siRNA, R2D2 was more efficiently photocrosslinked when the 5-iodouracil was on the strand more frequently incorporated into the RISC, whereas when the 5-iodouracil was on the strand less often incorporated into the RISC, Dcr-2 was more efficiently photocrosslinked (Fig. 2, A and B). Because a 5-iodouracil at p20 of one siRNA strand is near the 5′ end of the other strand, Dcr-2 must lie near the 5′ end of the strand entering the RISC (the guide strand), whereas

R2D2 binds near the 5′ end of the strand destined for destruction.

When we used the symmetric *sod1* siRNA (Fig. 2, C and D, siRNA *a*), Dcr-2 and R2D2 were photocrosslinked with nearly equal efficiency to the 5-iodouracil strand (Fig. 2D, siRNA *a*); this finding suggests that each protein binds about half the time to one or the other end of the siRNA. In contrast, when we used derivatives of this siRNA that contained single-nucleotide mismatches that made them highly asymmetric, the 5-iodouracil strand was photocrosslinked to either Dcr-2 or R2D2, but not to both (Fig. 2, C and D, siRNAs *b* and *c*). With the asymmetric siRNA sequence, the photocrosslinking data suggest that Dcr-2 is almost always near the 5′ end of the guide strand and R2D2 near the 5′ end of the passenger strand. As expected when both siRNA strands contained p20 5-iodouracil and 5′-[^{32}P]phosphate groups, both proteins were photocrosslinked (Fig. 2, C and D, siRNA *d*). When we used a reciprocal series of siRNAs in which the strands assembled into and excluded from the RISC were reversed (Fig. 2, C and D, siRNAs *e*, *f*, and *g*), Dcr-2 was again found near the 5′ end of the guide strand and R2D2 near the 5′ end of the passenger strand.

Purified, recombinant Dcr-2/R2D2 heterodimer alone can also sense the thermodynamic stabilities of the ends of an siRNA duplex. At physiologically relevant concentrations of the proteins (*16*), photocrosslinking reflected siRNA asymmetry (Fig. 3A). Like heterodimer binding to an siRNA (*17*), differential photocrosslinking of recombinant Dcr-2/R2D2 heterodimer to an siRNA (*18*) did not require adenosine triphosphate (ATP). In contrast, formation of the RLC requires ATP (*13*). The orientation of Dcr-2 and R2D2 on the siRNA duplex was less asymmetric for the recombinant heterodimer than for embryo lysate (compare Figs. 2D and 3A). We propose that siRNA asymmetry is initially sensed by the Dcr-2/R2D2 heterodimer in an ATP-independent manner but is later amplified by the ATP-dependent action of other proteins.

Photocrosslinking of R2D2, but not Dcr-2, to the two ends of an siRNA duplex was influenced by the presence of a 5′-phosphate group on the siRNA. We prepared a series of highly asymmetric siRNAs in which the strand containing the p20 5-iodouracil was radiolabeled with ^{32}P at the 5′ end and the other strand contained either a 5′-hydroxyl or 5′-phosphate group (Fig. 3B). In four trials, R2D2 photocrosslinking to the nearby p20 5-iodouracil of the guide strand was greater by a factor of 4.6 ± 0.4 (average ± SD) when the passenger strand contained a 5′-phosphate rather than a hydroxyl group (Fig. 3C, left, siRNAs *c* and *e*). R2D2 photocrosslinking in ATP-depleted embryo lysate likewise required a 5′-phosphate at the more thermodynamically stable siRNA end (Fig. 3C, right, siRNAs *c* and *e*). Thus, R2D2 can sense two aspects of siRNA structure: the stability of an siRNA 5′ end, and the presence of a 5′-phosphate group. In contrast, Dcr-2 photocrosslinking was unperturbed by a 5′-hydroxyl group on the guide strand, both for the purified protein and in ATP-depleted lysate (Fig. 3C, siRNAs *b* and *f*).

Active siRNAs contain 5′-phosphate groups on both strands (*3*, *11*, *19*–*21*). A 5′-phosphate on the guide strand is essential for siRNA function, but blocking 5′-phosphorylation of the passenger strand impairs rather than eliminates siRNA activity (*11*). Our results suggest a molecular explanation for this observation: A 5′-phosphate on the passenger strand enhances R2D2 binding, thereby facilitating efficient incorporation of an siRNA into the RLC and consequently into the RISC. Thus, R2D2 is a licensing factor that ensures that only authentic siRNAs enter the RNAi pathway in *Drosophila*.

Dcr-2 alone does not efficiently bind siRNA (*17*), nor can Dcr-2 alone be photocrosslinked to any of the siRNAs in this study (*18*). Taken together, these results and the data presented here suggest that orientation of the Dcr-2/R2D2 heterodimer is determined largely by R2D2 binding to the siRNA end with the most double-stranded character. This binding is presumably mediated by one or both of the R2D2 double-stranded RNA binding domains. A 5′ mismatch on an siRNA strand may therefore be an antideterminant for R2D2 binding, acting to direct the R2D2 protein to the 5′ end of the passenger

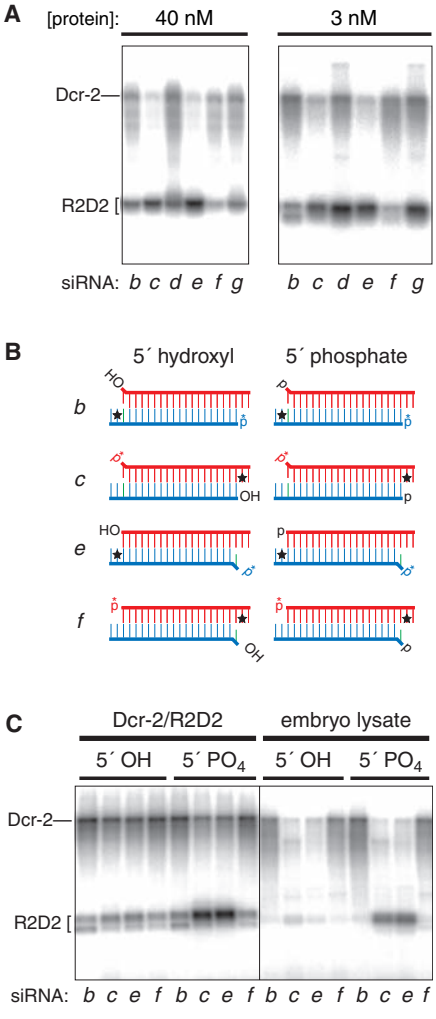


Fig. 3. The Dcr-2/R2D2 heterodimer alone can sense the asymmetry of an siRNA. (**A**) Photocrosslinking of recombinant Dcr-2/R2D2 heterodimer to the series of asymmetric siRNAs in Fig. 2C. (**B**) Structure of the siRNAs used in (C). The siRNAs all contained a single 5′-[^{32}P]phosphate on one strand and either a 5′-hydroxyl or 5′-phosphate group on the other. The siRNA sequences were as in Fig. 2C. (**C**) R2D2 senses the presence of a 5′-phosphate on the passenger strand. R2D2 photocrosslinking to the 5-iodouracil nearest the 5′ end of the passenger strand was reduced when the 5′ end of the passenger strand contained a 5′-hydroxyl rather than a 5′-phosphate group (siRNAs *c* and *e*); photocrosslinking of Dcr-2 was unaltered by the presence or absence of a 5′-phosphate on the guide strand (siRNAs *b* and *f*).

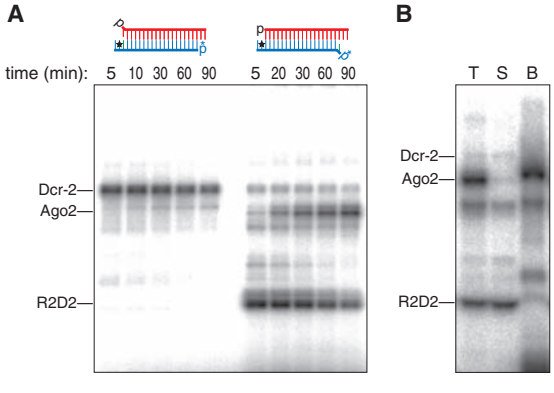


Fig. 4. Exchange of R2D2 for Ago2 at the 3′ end of the siRNA guide strand. (**A**) ^{32}P-radiolabeled siRNAs *b* and *e* (Fig. 2B) were incubated with embryo lysate in the presence of ATP for the times indicated, and binding of proteins near the 3′ end of the siRNA guide and passenger strands was monitored by photocrosslinking. Indicated times include the 4 min during which the sample was exposed to ultraviolet irradiation at room temperature. (**B**) ^{32}P-radiolabeled siRNA *c* was incubated with embryo lysate and photocrosslinked. The photocrosslinked proteins were then captured with a 2′-*O*-methyl oligonucleotide complementary to the siRNA guide strand. T, total reaction before incubation with the tethered oligonucleotide; S, supernatant after incubation; B, siRNA-photocrosslinked proteins bound to the tethered oligonucleotide.

strand and positioning Dcr-2 near the 5′ end of the strand to be loaded into the RISC. In this model, R2D2, as a component of the Dcr-2/R2D2 heterodimer, is the primary protein sensor of siRNA thermodynamic asymmetry.

How does the RLC, with the Dcr-2/R2D2 heterodimer positioned asymmetrically on the siRNA, progress to the RISC? Argonaute 2 (Ago2) is a ~130-kD protein that is a core component of the RISC (*22*) and is required for siRNA unwinding (*14*). We found that a ~130-kD protein was crosslinked to siRNA when the guide strand contained 5-iodouracil at p20 (asterisk in Fig. 2C, siRNAs *c*, *d*, *e*, and *g*). The ~130-kD protein was photocrosslinked only to the guide strand of the siRNA (Fig. 4), which suggests that this protein is a component of the RISC. The ~130-kD protein was immunoprecipitated with antibodies to Ago2 but not to Ago1 (fig. S3A) and was not observed in embryos lacking both maternal and zygotic Ago2 (*ago2*^{414}, fig. S3B). Thus, the ~130-kD protein is Ago2. When R2D2 and Ago2 were photocrosslinked to siRNAs *b* or *e* (which contain 5-iodouracil at p20 of the passenger or the guide strand), R2D2 was bound to the 3′ end of the guide strand and Dcr-2 to the 3′ end of the passenger strand at early times in the reaction (Fig. 4A). Later, binding of R2D2 and Dcr-2 decreased concurrently, accompanied by a corresponding increase in binding of Ago2 to the 3′ end of the guide strand. In *ago2*^{414} lysates, R2D2 binding to the 3′ end of the guide strand and Dcr-2 binding to the 3′ end of the passenger strand did not decrease with time (fig. S4A); this finding suggests that binding of Ago2 facilitates the release of the heterodimer from siRNA.

The siRNA bound by Ago2 is single-stranded, because Ago2, when photocrosslinked to siRNA, was captured by a tethered 2′-*O*-methyl oligonucleotide complementary to the siRNA guide strand (Fig. 4B) (*23*), as has been observed for the RISC (*7*, *23*–*25*). R2D2 was not captured by the 2′-*O*-methyl oligonucleotide, but was instead recovered in the supernatant, consistent with R2D2 binding of double-stranded siRNA.

Our data suggest a model for RISC assembly. First, R2D2 orients the Dcr-2/R2D2 heterodimer on the siRNA within the RLC. As siRNA unwinding proceeds, the heterodimer is exchanged for Ago2, the core component of the RISC. Indeed, we cannot detect single-stranded siRNA in the RLC assembled in *ago2*^{414} lysate (fig. S4, B and C). We hypothesize that unwinding occurs only when Ago2 is available, so that siRNA in the RLC is unwound only when the RISC can be assembled.

References and Notes

1. A. J. Hamilton, D. C. Baulcombe, *Science* **286**, 950 (1999).
2. S. M. Hammond, E. Bernstein, D. Beach, G. J. Hannon, *Nature* **404**, 293 (2000).
3. S. M. Elbashir, W. Lendeckel, T. Tuschl, *Genes Dev.* **15**, 188 (2001).
4. J. Martinez, T. Tuschl, *Genes Dev.* **18**, 975 (2004).
5. D. S. Schwarz, Y. Tomari, P. D. Zamore, *Curr. Biol.* **14**, 787 (2004).
6. P. D. Zamore, T. Tuschl, P. A. Sharp, D. P. Bartel, *Cell* **101**, 25 (2000).
7. D. S. Schwarz *et al.*, *Cell* **115**, 199 (2003).
8. A. Khvorova, A. Reynolds, S. D. Jayasena, *Cell* **115**, 209 (2003).
9. P. Aza-Blanc *et al.*, *Mol. Cell* **12**, 627 (2003).
10. Y. S. Lee *et al.*, *Cell* **117**, 69 (2004).
11. A. Nykänen, B. Haley, P. D. Zamore, *Cell* **107**, 309 (2001).
12. J. W. Pham, J. L. Pellino, Y. S. Lee, R. W. Carthew, E. J. Sontheimer, *Cell* **117**, 83 (2004).
13. Y. Tomari *et al.*, *Cell* **116**, 831 (2004).
14. K. Okamura, A. Ishizuka, H. Siomi, M. C. Siomi, *Genes Dev.* **18**, 1655 (2004).
15. N. Doi *et al.*, *Curr. Biol.* **13**, 41 (2003).
16. See supporting data on *Science* Online.
17. Q. Liu *et al.*, *Science* **301**, 1921 (2003).
18. Y. Tomari, C. Matranga, B. Haley, N. Martinez, P. D. Zamore, data not shown.
19. Y.-L. Chiu, T. M. Rana, *Mol. Cell* **10**, 549 (2002).
20. A. Boutla, C. Delidakis, I. Livadaras, M. Tsagris, M. Tabler, *Curr. Biol.* **11**, 1776 (2001).
21. D. S. Schwarz, G. Hutvágner, B. Haley, P. D. Zamore, *Mol. Cell* **10**, 537 (2002).
22. S. M. Hammond, S. Boettcher, A. A. Caudy, R. Kobayashi, G. J. Hannon, *Science* **293**, 1146 (2001).
23. G. Hutvágner, M. J. Simard, C. C. Mello, P. D. Zamore, *PLoS Biol.* **2**, 465 (2004).
24. G. Meister, M. Landthaler, Y. Dorsett, T. Tuschl, *RNA* **10**, 544 (2004).
25. B. Haley, P. D. Zamore, *Nature Struct. Mol. Biol.* **11**, 599 (2004).

26. We thank D. Turner, C. R. Matthews, Z. Gu, and members of the Zamore laboratory for advice and support, and Q. Liu, G. Hannon, T. Uemura, D. Smith, R. Carthew, M. Siomi, and H. Siomi for gifts of reagents. Y.T. is a recipient of a long-term fellowship from the Human Frontier Science Program. P.D.Z. is a Pew Scholar in the Biomedical Sciences and a W. M. Keck Foundation Young Scholar in Medical Research. Supported by NIH grants GM62862-01 and GM65236-01 (P.D.Z.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1377/DC1
Materials and Methods
Table S1
Figs. S1 to S4
References

14 July 2004; accepted 20 September 2004

The Human Polyomavirus, JCV, Uses Serotonin Receptors to Infect Cells

Gwendolyn F. Elphick,[1] William Querbes,[1,2] Joslynn A. Jordan,[1,2] Gretchen V. Gee,[1,3] Sylvia Eash,[1,2] Kate Manley,[1,3] Aisling Dugan,[1,2] Megan Stanifer,[1,3] Anushree Bhatnagar,[4] Wesley K. Kroeze,[4] Bryan L. Roth,[4] Walter J. Atwood[1,2,3]*

The human polyomavirus, JCV, causes the fatal demyelinating disease progressive multifocal leukoencephalopathy in immunocompromised patients. We found that the serotonergic receptor $5HT_{2A}$R could act as the cellular receptor for JCV on human glial cells. The $5HT_{2A}$ receptor antagonists inhibited JCV infection, and monoclonal antibodies directed at $5HT_{2A}$ receptors blocked infection of glial cells by JCV, but not by SV40. Transfection of $5HT_{2A}$ receptor–negative HeLa cells with a $5HT_{2A}$ receptor rescued virus infection, and this infection was blocked by antibody to the $5HT_{2A}$ receptor. A tagged $5HT_{2A}$ receptor colocalized with labeled JCV in an endosomal compartment following internalization. Serotonin receptor antagonists may thus be useful in the treatment of progressive multifocal leukoencephalopathy.

The incidence of progressive multifocal leukoencephalopathy (PML) has increased 50-fold since 1979 and now affects nearly 1 in every 200,000 persons (*1*). The disease is due to infection of oligodendrocytes by the common human polyomavirus, JCV (*2*). Initial infection with JCV occurs early in childhood and eventually reaches a seroprevalence of between 70 and 80% in the adult population. The initial infection is subclinical, and the virus establishes a lifelong persistent infection. At any given time, ~5% of the population is actively excreting virus in the urine, and JCV is a frequent contaminant of untreated human sewage (*3*). PML occurs almost exclusively in severely immunosuppressed patients. The majority of cases occur in patients with AIDS, and to date there is no effective treatment (*4*). PML is initiated when JCV traffics from peripheral sites, such as the kidney and lymphoid organs, to the central nervous system (CNS) by unknown mechanisms. There is a strong association between JCV and human B lymphocytes, and the virus may traffic to the CNS in an

[1]Department of Molecular Microbiology and Immunology, [2]Graduate Program in Pathobiology, [3]Graduate Program in Molecular Biology, Cellular Biology, and Biochemistry, Brown University, Providence, RI 02912, USA. [4]Department of Biochemistry, Case Western Reserve University Medical School, Cleveland, Ohio 44106, USA.

*To whom correspondence should be addressed. E-mail: Walter_Atwood@Brown.edu

infected B cell (*5*–*7*). Once in the CNS, JCV infects both oligodendrocytes and astrocytes.

N-linked glycoproteins containing terminal alpha 2-6–linked sialic acid are a critical component of the JCV receptor (*8*). The tissue distribution of this receptor-type sialic acid strongly correlates with the known tropism of JCV for oligodendrocytes, astrocytes, B-lymphocytes, and kidney epithelial cells (*9*). JCV receptor interactions play a critical role in tropism, because a hybrid SV40 virus containing JCV capsid proteins maintains the restricted host range of JCV (*10*). Also, unlike the related polyomavirus SV40, which enters cells by caveolae-dependent endocytosis, JCV enters cells by a ligand-inducible clathrin-dependent pathway (*11*–*15*).

Chlorpromazine, which blocks clathrin-dependent endocytosis, and the related compound clozapine effectively block JCV infection of glial cells (*16*). Both chlorpromazine and clozapine belong to a class of drugs known as serotonin-dopamine inhibitors (SDIs). Because glial cells express receptors for both dopamine and serotonin (Fig. 1A), we hypothesized that JCV may use either serotonin receptors or dopamine receptors to infect glial cells. To test this hypothesis, glial cells were treated with increasing concentrations of the dopamine antagonists bromocriptine and minaprine, and a dopamine agonist, pergolide. These agents have generally minimal activity against serotonin receptors (*17*, *18*). Glial cells were also treated with increasing concentrations of antagonists with activity against both dopamine and serotonin receptors (*19*, *20*). These included metoclopramide, chlorpromazine, and clozapine. The cells were then incubated with JCV at a multiplicity of infection (MOI) of 1.0 in the continued presence of drug. At 72 hours after infection, the cells were assayed for viral infection. The dopamine-specific antagonists, bromocriptine and minaprine, and the dopamine agonist, pergolide, had little to no effect on the infectivity of glial cells by JCV (Fig. 1B). In contrast, metoclopramide, chlorpromazine, and clozapine, which antagonize the 5HT2 serotonergic receptors, all significantly inhibited infection (Fig. 1B). Because these reagents are not highly specific, we next asked whether 5HT itself or selective 5HT2 receptor antagonists could inhibit JCV infection. Glial cells were treated in triplicate with increasing concentrations of 5HT (which down-regulates serotonin receptors), MDL100.907 (which selectively inhibits $5HT_{2A}R$), SB206553 (which inhibits $5HT_{2C}$ R), ketanserin (which inhibits $5HT_{2A}R$ and $5HT_{2C}$ R), or ritanserin (which inhibits $5HT_{2A}R$, $5HT_{2B}R$, and $5HT_{2C}R$) (*21*–*25*). 5HT and MDL100.907 both inhibited infection of glial cells by JCV at concentrations of 1.0 μM (Fig. 1B). The $5HT_{2C}$ inhibitor SB206553 only slightly inhibited infection when used at 1.0 μM (Fig. 1B). The $5HT_{2A}$ and $5HT_{2C}$ inhibitor ketanserin inhibited infection at 0.1 μM, and ritanserin also inhibited at 0.1 μM (Fig. 1B).

We next asked whether antibodies directed at $5HT_{2A}R$, $5HT_{2C}$ R, or at the D1, D2, and D3 dopamine receptors could block in-

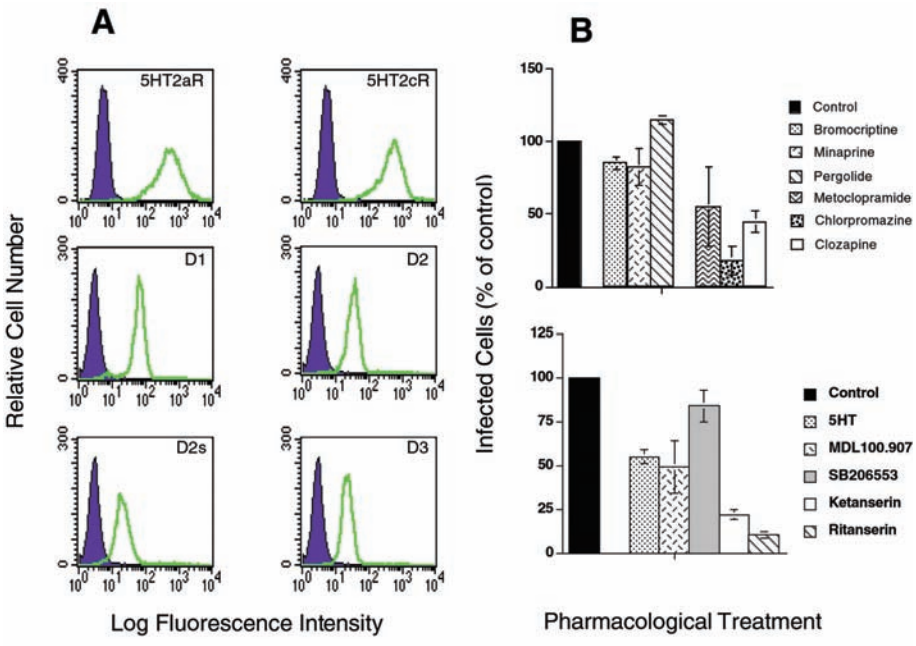


Fig. 1. (**A**) Glial cells express receptors for both serotonin and dopamine. Glial cells were incubated with irrelevant antibody (solid histograms), with monoclonal antibodies to the $5HT_{2A}$ and $5HT_{2C}$ serotonergic receptors (top panels, open histograms), or with polyoclonal antibodies to the D1, D2, D2s, and D3 dopamine receptors (middle and bottom panels, open histograms). Antibody binding was detected with either goat anti-mouse or goat anti-rabbit secondary antibodies conjugated to AlexaFluor-488. (**B**) SDIs inhibited infection of glial cells by JCV. Glial cells were incubated with the dopamine antagonists bromocriptine and minaprine, a dopamine agonist, pergolide, or with metoclopramide, chlorpromazine, or clozapine, which antagonize both dopamine receptors and serotonin receptors. Cells were then challenged with JCV and infection scored at 72 hours after infection by indirect immunofluorescence assay of V antigen–expressing cells. The percentage of infected cells in untreated cultures was set at 100%. The ability of these agents to inhibit JCV infection correlate with their ability to antagonize $5HT_{2A}$ and $5HT_{2c}$ serotonin receptors. (Bottom panel) 5HT and specific 5HT2 antagonists inhibited infection of glial cells by JCV. Glial cells were incubated with 5HT, MDL100.907, SB206553, ketanserin, or ritanserin. Cells were challenged and scored for viral infection as described above. 5HT, MDL100.907, ketanserin, and ritanserin all significantly inhibited infection. SB206553 had a modest inhibitory effect on JCV infection.

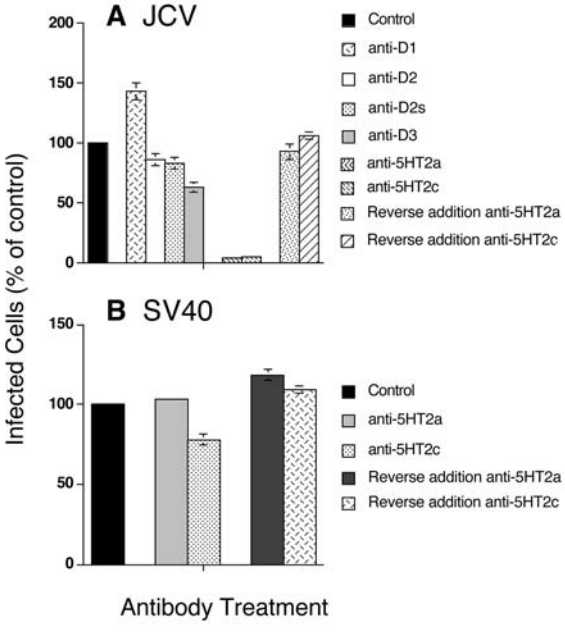


Fig. 2. (**A**) Antibodies directed at $5HT_{2A}R$ or $5HT_{2C}$ R, but not antibodies directed at dopamine receptors, inhibit JCV infection of glial cells. Glial cells were incubated with antibodies against either serotonin receptors or dopamine receptors as indicated. The cells were then challenged with JCV and infection scored at 72 hours after infection by indirect immunofluorescence analysis of V antigen–expressing cells. The percentage of infected cells in untreated cultures was set at 100%. If antibodies to $5HT_{2A}R$ or $5HT_{2C}$ R were added at 24 hours after infection, they had no effect on virus infection. (**B**) Antibodies directed at $5HT_{2A}R$ or $5HT_{2C}$ R do not inhibit infection of glial cells by SV40, as indicated.

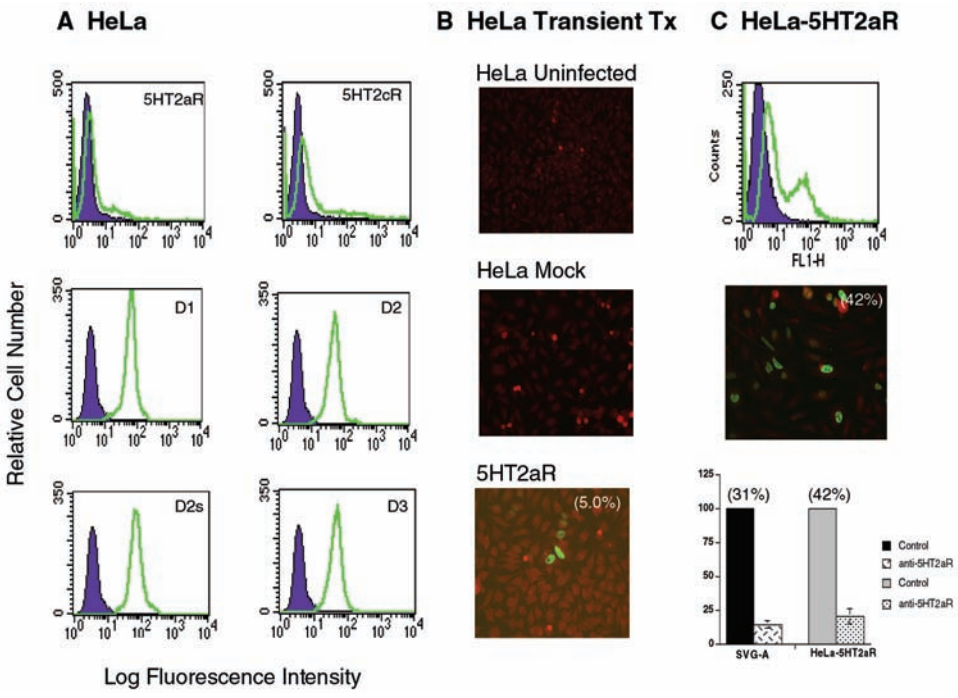


Fig. 3. (**A**) HeLa cells express dopamine receptors but do not express $5HT_{2A}$ or $5HT_{2C}$ serotonergic receptors. HeLa cells were incubated with irrelevant antibody (solid histograms), with monoclonal antibodies to the $5HT_{2A}$R and $5HT_{2C}$ R (top panels, open histograms), or with polyclonal antibodies to the D1, D2, D2s, and D3 dopamine receptors (middle and bottom panels, open histograms). Antibody binding was detected with either goat anti-mouse or goat anti-rabbit secondary antibodies conjugated to AlexaFluor-488. (**B**) Transient transfection of HeLa cells with the $5HT_{2A}$ receptor rescues virus infection. HeLa cells were untreated, transfected with irrelevant plasmid DNA (mock), or with a $5HT_{2A}$ receptor–expressing plasmid. At 24 hours after transfection, the cells were challenged with JCV and infection scored 48 hours later by indirect immunofluorescent analysis of T antigen–expressing cells. The cells were counterstained with Evans blue, which fluoresces red in the ultraviolet channel. T antigen–expressing cells could only be detected in HeLa cells transfected with the $5HT_{2A}$ receptor clone. The percentage of T antigen–positive cells is indicated in parentheses. (**C**) Stable transfection of HeLa cells with the $5HT_{2A}$ receptor rescues virus infection. (Top panel) Analysis of $5HT_{2A}$R expression on HeLa cells stably expressing $5HT_{2A}$R (HeLa-$5HT_{2A}$R). (Middle panel) HeLa-$5HT_{2A}$R cells infected with JCV and stained for T antigen. The percentage of infected cells is indicated in parentheses. (Bottom panel) Infection of SVG-A and HeLa-$5HT_{2A}$R cells by JCV is blocked by antibodies to $5HT_{2A}$R. The percentage of infected SVG-A and HeLa-$5HT_{2A}$R cells in nonspecific antibody–treated controls is indicated in parentheses. The amount of infection in the control samples was set at 100% for comparison.

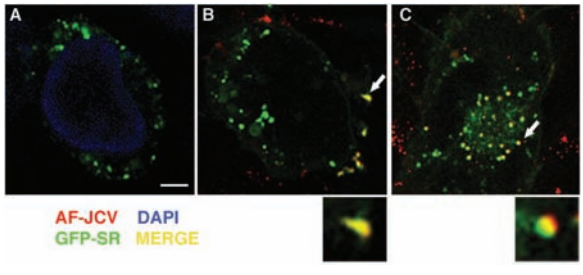


Fig. 4. Colocalization of JCV and $5HT_{2A}$ receptor at different stages of virus internalization. (**A**) Glial cells were transiently transfected with a $5HT_{2A}$R-GFP serotonin receptor construct, and the fusion protein was visualized by confocal microscopy. (**B** and **C**) $5HT_{2A}$R-GFP–expressing cells were exposed to AlexaFluor-594–labeled JCV (red) and allowed to internalize for 5 min (B) or 30 min (C). At both time points, virus is shown to colocalize with the $5HT_{2A}$-GFP receptor (yellow). Insets show an enlarged portion of the image identified by arrows. Scale bar, 10 μ.

fection. Glial cells were pre-incubated with equivalent amounts of each antibody and then challenged with JCV. None of the antibodies to dopamine receptors specifically inhibited infection of glial cells by JCV (Fig. 2A). In contrast, the antibodies to both $5HT_{2A}$R and $5HT_{2C}$ R significantly inhibited infection (Fig. 2A). These antibodies had no effect on infection if added 24 hours after infection (Fig. 2A). As a control for specificity, we pre-incubated glial cells with antibodies to $5HT_{2A}$R or $5HT_{2C}$ R and then challenged with the related virus, SV40. These antibodies had no significant effect on infection of glial cells by SV40 (Fig. 2B).

We next asked if we could rescue infection in $5HT_{2A}$ receptor–negative cells by transient and/or stable expression of a $5HT_{2A}$ receptor clone. HeLa cells did not express either $5HT_{2A}$ or $5HT_{2C}$ receptors but expressed abundant levels of D1, D2, D2s, and D3 dopamine receptors (Fig. 3A). We transiently transfected HeLa cells with p$5HT_{2A}$R or a control vector, and at 24 hours after transfection the cells were infected with JCV at an MOI of 10.0. Infection was assayed 48 hours after infection. HeLa cells transfected with the control construct (mock) remained refractory to JCV infection (Fig. 3B). In contrast, HeLa cells transfected with the $5HT_{2A}$ receptor clone became susceptible to infection (Fig. 3B). The percentage of infected cells was low (5%) but consistent with the low transfection efficiency of HeLa cells. We next established a HeLa cell line stably expressing the $5HT_{2A}$ receptor by cotransfection of p$5HT_{2A}$R with a plasmid encoding resistance to puromycin (pMSCVpuro) (Fig. 3C). HeLa-$5HT_{2A}$R cells were then challenged with JCV in the presence and absence of antibodies to $5HT_{2A}$R. The HeLa-$5HT_{2A}$R cells were readily infected by JCV at levels comparable to infection in SVG-A glial cells (Fig. 3C). Infection of both cell types by JCV was blocked by antibody to $5HT_{2A}$R (Fig. 3C).

Glial cells were transfected with a GFP-tagged $5HT_{2A}$ receptor clone and then incubated with Alexa-fluor 594–labeled JCV at 24 hours after transfection when GFP expression was maximal (*26*). Virus binding was first synchronized by incubation with the cells at 4°C for 30 min. The cells were then either fixed immediately or warmed to 37°C for 5 min or 30 min and then fixed. When the cells were allowed to internalize virus at 37°C for 5 min or 30 min, strong colocalization between the virus and the $5HT_{2A}$ receptors was seen (Fig. 4). The virus appeared to initially interact only with the alpha 2-6–linked sialic acid component of the JCV receptor, and then at 37°C interacted with the $5HT_{2A}$ receptor. This second interaction most likely leads to efficient and rapid virus internalization. This is not unexpected, because both JCV and $5HT_{2A}$ receptors are rapidly internalized by clathrin-dependent endocytosis after ligand binding. This is also consistent with the fact that JCV internalization is accompanied by activation of the MAP kinases ERKs1 and 2, because serotonin binding to $5HT_{2A}$ receptors also activates ERKs 1 and 2 (*12*, *27*, *28*).

Compared with other polyomaviruses, JCV has a very restricted tropism, infecting

oligodendrocytes, astrocytes, kidney epithelial cells, and, to a limited extent, B lymphocytes. In vitro, the virus can only be efficiently propagated in primary human fetal glial cells or in human fetal glial cell lines such as POJ and SVG (*29*–*31*). This restricted tropism is due to the presence or absence of cell-type-specific transcription and replication factors and to the presence of specific virus receptors. HeLa cells are completely refractory to infection by JCV but will support early viral gene expression when transfected with JCV DNA. HeLa cells express the JCV receptor–type sialic acid (α 2-6 SA) and bind virus as well as permissive glial cells, suggesting that sialic acid is not sufficient for mediating virus infection (*32*). Our ability to rescue JCV infection in receptor-negative HeLa cells by transiently or stably introducing the $5HT_{2A}$ receptor demonstrates that $5HT_{2A}R$ is a functional entry receptor for JCV. The breadth of other serotonergic receptors that might also function as JCV receptors has not been thoroughly investigated, but preliminary data have ruled out the 5HT1, 5HT3, and 5HT7 families.

Neurons express abundant levels of serotonin receptors but are generally refractory to infection by JCV. However, neurons do not express the receptor-type sialic acid for JCV, which indicates that infection of cells requires both components of the JCV receptor (*9*). Oligodendrocytes, astrocytes, B lymphocytes, and kidney epithelial cells all express both the alpha 2-6–linked sialic acid component of the JCV receptor and $5HT_{2A}$ receptors (*9*, *33*–*39*).

5HT2-family receptors are highly expressed on brain microvasculature, on astrocytes at the blood-brain barrier, and in brain regions lacking the blood-brain barrier, such as the area postrema and the choroid plexus. This raises the possibility that JCV may directly traffic to the CNS via the blood under viremic conditions, as occurs during severe and prolonged immunosuppression.

Finally, serotonin receptor agonists and antagonists are widely used to treat a variety of neurological and psychiatric disorders. Drugs that have been developed to treat PML have all been hampered by poor bioavailability in the CNS, a problem not inherent to serotonergic inhibitors. Prophylactic treatment of HIV-infected patients with serotonergic antagonists may prevent the spread of JCV to the CNS and the development of PML. Aggressive therapeutic treatment of patients with PML may reduce viral spread within the CNS and prevent additional episodes of demyelination.

References and Notes

1. R. C. Holman, T. J. Torok, E. D. Belay, R. S. Janssen, L. B. Schonberger, *Neuroepidemiology* **17**, 303 (1998).
2. D. L. Walker, R. J. Frisque, in *The Papovaviridae*, N. P. Salzman, ed. (Plenum, New York and London, 1986), pp. 327–377.
3. S. Bofill-Mas, S. Pina, R. Girones, *Appl. Environ. Microbiol.* **66**, 238 (2000).
4. J. R. Berger, E. O. Major, *Semin. Neurol.* **19**, 193 (1999).
5. S. A. Houff *et al.*, *N. Engl. J. Med.* **318**, 301 (1988).
6. M. G. C. Monaco, W. J. Atwood, M. Gravell, C. S. Tornatore, E. O. Major, *J. Virol.* **70**, 7004 (1996).
7. M. C. Monaco, P. N. Jensen, J. Hou, L. C. Durham, E. O. Major, *J. Virol.* **72**, 9918 (1998).
8. C. K. Liu, G. Wei, W. J. Atwood, *J. Virol.* **72**, 4643 (1998).
9. S. Eash *et al.*, *Am. J. Pathol.* **164**, 419 (2004).
10. B. J. Chen, W. J. Atwood, *Virology* **300**, 282 (2002).
11. M. T. Pho, A. Ashok, W. J. Atwood, *J. Virol.* **74**, 2288 (2000).
12. W. Querbes, A. Benmerah, D. Tosoni, P. P. Di Fiore, W. J. Atwood, *J. Virol.* **78**, 250 (2004).
13. L. Pelkmans, J. Kartenbeck, A. Helenius, *Nature Cell Biol.* **3**, 473 (2001).
14. L. C. Norkin, *Immunol. Rev.* **168**, 13 (1999).
15. L. C. Norkin, H. A. Anderson, S. A. Wolfrom, A. Oppenheim, *J. Virol.* **76**, 5156 (2002).
16. S. Baum *et al.*, *J. Neurovirol.* **9**, 32 (2003).
17. A. Newman-Tancredi *et al.*, *J. Pharmacol. Exp. Ther.* **303**, 815 (2002).
18. M. Velasco, A. Luchsinger, *Am. J. Ther.* **5**, 37 (1998).
19. K. Herrick-Davis, E. Grinde, M. Teitler, *J. Pharmacol. Exp. Ther.* **295**, 226 (2000).
20. P. K. Gillman, *J. Psychopharmacol.* **13**, 100 (1999).
21. F. G. Boess, I. L. Martin, *Neuropharmacology* **33**, 275 (1994).
22. M. S. Choudhary, S. Craigo, B. L. Roth, *Mol. Pharmacol.* **42**, 627 (1992).
23. M. Dudley, A. Ogden, A. Carr, T. Nieduzak, J. Kehne, *Soc. Neurosci. Abstr.* **16**, 1037 (1990).
24. J. L. Herndon, A. Ismaiel, S. P. Ingher, M. Teitler, R. A. Glennon, *J. Med. Chem.* **35**, 4903 (1992).
25. K. Kristiansen, S. G. Dahl, *Eur. J. Pharmacol.* **306**, 195 (1996).
26. A. Bhatnagar *et al.*, *J. Biol. Chem.* **276**, 8269 (2001).
27. D. Hoyer *et al.*, *Pharmacol. Rev.* **46**, 157 (1994).
28. A. Bhatnagar, D. J. Sheffler, W. K. Kroeze, B. Compton-Toth, B. L. Roth, *J. Biol. Chem.* **279**, 34614 (2004).
29. C. Mandl, D. L. Walker, R. J. Frisque, *J. Virol.* **61**, 755 (1987).
30. B. Padgett, G. ZuRhein, D. Walker, R. Echroade, B. Dessel, *Lancet* **I**, 1257 (1971).
31. E. O. Major *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1257 (1985).
32. G. Wei, C. K. Liu, W. J. Atwood, *J. Neurovirol.* **6**, 127 (2000).
33. A. Merzak, S. Koochekpour, M. P. Fillion, G. Fillion, G. J. Pilkington, *Brain Res. Mol. Brain Res.* **41**, 1 (1996).
34. Z. Cohen *et al.*, *J. Cereb. Blood Flow Metab.* **19**, 908 (1999).
35. M. I. Fonseca, Y. G. Ni, D. D. Dunning, R. Miledi, *Brain Res. Mol. Brain Res.* **89**, 11 (2001).
36. S. Belachew *et al.*, *Neuroreport* **9**, 973 (1998).
37. J. A. Gray *et al.*, *Mol. Pharmacol.* **60**, 1020 (2001).
38. D. W. Bonhaus *et al.*, *Br. J. Pharmacol.* **115**, 622 (1995).
39. M. Chang, L. Zhang, J. P. Tam, E. Sanders-Bush, *J. Biol. Chem.* **275**, 7021 (2000).

40. We would like to thank all members of the Atwood laboratory for critical discussion during the course of this work. We thank J. Sedivy for critical discussions during the preparation of the manuscript. We also thank L. Brossay for the pMSCV plasmid, R. Creton for critical help with confocal microscopy, and A. Robinson, A. Bozek, and L. St. Pierre for administrative assistance. Work in our laboratory was supported by a grant from the National Cancer Institute, R01 CA71878, and by a grant from the National Institute of Neurological Disorders and Stroke, R01 NS43097. W.Q. is supported by a Graduate Assistantship in Areas of National Need training grant from the Department of Education, P200A030100. Work in the Roth lab was supported by RO1MH57635, RO1MH61887, and the National Institute of Mental Health Psychoactive Drug Screening Program to B.L.R.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1380/DC1
Materials and Methods

2 August 2004; accepted 21 September 2004

Fat Mobilization in Adipose Tissue Is Promoted by Adipose Triglyceride Lipase

Robert Zimmermann,[1]* Juliane G. Strauss,[1]* Guenter Haemmerle,[1] Gabriele Schoiswohl,[1] Ruth Birner-Gruenberger,[3] Monika Riederer,[1] Achim Lass,[1] Georg Neuberger,[2] Frank Eisenhaber,[2] Albin Hermetter,[3] Rudolf Zechner[1]†

Mobilization of fatty acids from triglyceride stores in adipose tissue requires lipolytic enzymes. Dysfunctional lipolysis affects energy homeostasis and may contribute to the pathogenesis of obesity and insulin resistance. Until now, hormone-sensitive lipase (HSL) was the only enzyme known to hydrolyze triglycerides in mammalian adipose tissue. Here, we report that a second enzyme, adipose triglyceride lipase (ATGL), catalyzes the initial step in triglyceride hydrolysis. It is interesting that ATGL contains a "patatin domain" common to plant acyl-hydrolases. ATGL is highly expressed in adipose tissue of mice and humans. It exhibits high substrate specificity for triacylglycerol and is associated with lipid droplets. Inhibition of ATGL markedly decreases total adipose acyl-hydrolase activity. Thus, ATGL and HSL coordinately catabolize stored triglycerides in adipose tissue of mammals.

Animals, seed plants, and fungi commonly store excessive amounts of energy substrates in the form of intracellular triglyceride (TG) deposits. In mammals, TGs are stored in adipose tissue, where they provide the primary source of energy during peri-

ods of food deprivation. Whole-body energy homeostasis depends on the precisely regulated balance of lipid storage and mobilization. Mobilization of stored fat is mediated by lipolytic enzymes, which degrade adipose TGs and release nonesterified fatty acids (FAs) into the circulation. Dysregulation of TG-lipolysis has been linked to variation in the concentration of circulating FA, an established risk factor for the development of insulin resistance in type 2 diabetes and related disorders (*1–4*).

During periods of increased energy demand, lipolysis in adipocytes is activated by hormones, such as catecholamines. Hormone interaction with G protein–coupled receptors results in increased adenylate cyclase activity, increased adenosine 3′,5′-monophosphate (cAMP) levels, and the activation of cAMP-dependent protein kinase (protein kinase A, PKA) (*5*). PKA phosphorylates two important proteins with established functions in lipolysis: HSL, an enzyme that catabolizes adipose tissue TGs, and perilipin A, an abundant structural protein located on the surface of lipid droplets. These modifications induce the translocation of HSL from the cytoplasm to the lipid droplet, where efficient TG hydrolysis occurs (*6*). Current models depict HSL as the rate-limiting enzyme in TG mobilization. However, the nonobese phenotype of HSL knock-out (HSL-KO) mice (*7–9*) and the accumulation of diglycerides (DGs) in their adipose tissue (*10*) suggest that there may be one or more additional lipases in adipose tissue that preferentially hydrolyzes the first ester bond of the TG molecule.

To search for such TG lipases, we screened gene and protein databases for murine and human proteins with structural homologies to known lipases, i.e., the GXSXG motif for serine esterases and α/β hydrolase folds. Candidates were analyzed for TG-hydrolase activity and expression in mouse adipose tissue. Only one previously undescribed enzyme fulfilled these requirements, and we named it "adipose triglyceride lipase" (ATGL).

The murine gene for ATGL (NCBI nucleotide entry AK031609) encodes a 486–amino acid protein (BAC27476) with a calculated molecular mass of 54 kD. The amino acid sequences of murine ATGL and two closely related proteins, NP_473429 (annotated as adiponutrin) and XP_128189 are shown in fig. S1. The human ATGL gene, also designated TTS-2.2, encodes a 504–amino acid protein (NP_065109) with 86% identity to the mouse enzyme. The N-terminal regions of ~260 residues in both the murine and the human enzyme contain a "predicted esterase of the α/β hydrolase fold" domain (COG1752) (*11*), as well as a GXSXG site with a putative active serine (amino acid 47). Moreover, a "patatin" domain (Pfam01734) can be detected in the same region (*12*). Patatin domain–containing proteins are commonly found in plant storage proteins such as the prototype patatin, an abundant protein of potato tubers (*13*). These proteins have been shown to have acyl-hydrolase activity on phospholipid, monoglyceride, and DG substrates. Patatin-domains are also present in TGL3, a TG-lipase of *Saccharomyces cerevisiae* (*14*), and human cytosolic phospholipase A2 (*15*).

For ATGL, mRNA is expressed at high levels in murine white and brown adipose tissue (WAT and BAT, Fig. 1A) and to a lesser degree in testis, cardiac muscle, and skeletal muscle. Highest expression of human ATGL (TTS-2.2) mRNA is also found in adipose tissue (fig. S2). ATGL mRNA expression was first detected 4 days after induction of differentiation of murine 3T3-L1 adipocytes, and maximum expression was observed at day 6 (Fig. 1B). To investigate whether ATGL hydrolyzes neutral lipids, we transfected simian

[1]Institute of Molecular Biosciences, University of Graz, Graz, Austria. [2]Research Institute of Molecular Pathology, Vienna, Austria. [3]Institute of Biochemistry, Graz University of Technology, Austria.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: rudolf.zechner@uni-graz.at

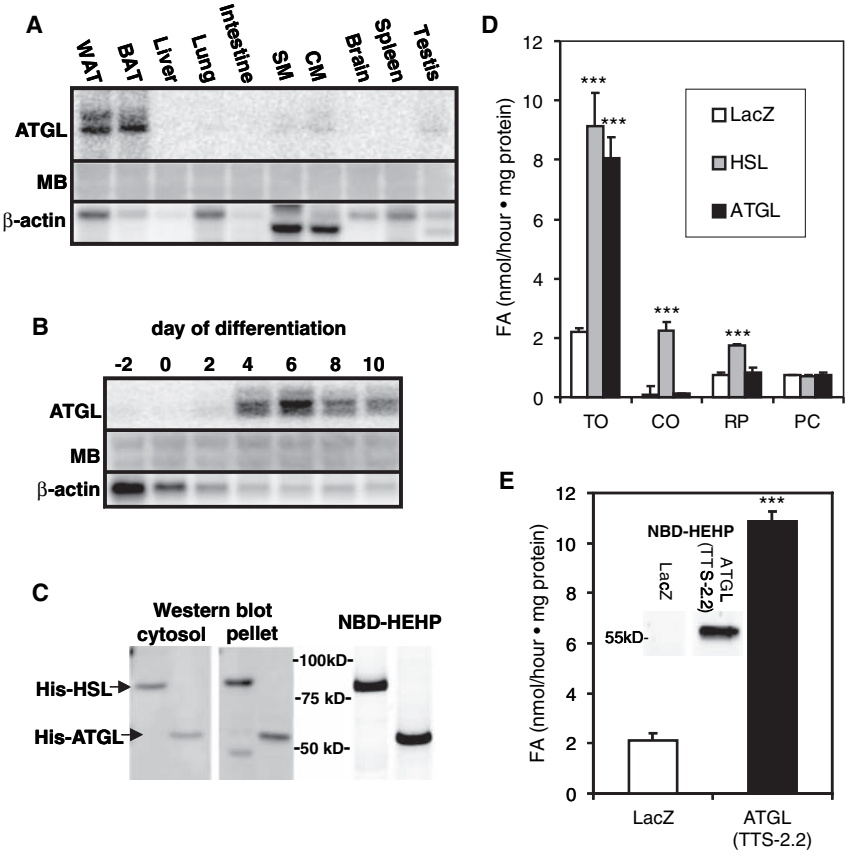


Fig. 1. Analysis of ATGL gene expression and enzyme activity. Total RNA (10 μg) from tissues of fasted mice (**A**) and 3T3-L1 adipocytes at various stages of differentiation (**B**) were subjected to Northern blot analysis. Specific mRNAs were detected with a radiolabeled mouse ATGL cDNA probe. Blots were stained for RNA by methylene blue (MB). A radiolabeled β-actin cDNA probe served as positive hybridization control. Abbreviations: WAT, white adipose tissue; BAT, brown adipose tissue; SM, skeletal muscle; CM, cardiac muscle. (**C**) His-tagged murine HSL and ATGL were detected by Western blotting in cytosolic extracts (100,000*g* supernatant) and membrane fractions (100,000*g* pellet) of transiently transfected COS-7 cells by using a monoclonal antibody against His. In addition, His-tagged murine HSL and ATGL were analyzed by binding of the fluorescent lipase inhibitor NBD-HEHP. Cytoplasmic extracts were preincubated with NBD-HEHP, subjected to SDS-PAGE. NBD-HEHP-labeled proteins were visualized by a BioRad FX Pro Laserscanner (*17*). (**D**). Lipid-hydrolase assays of cytosolic extracts of COS-7 cells expressing murine His-tagged ATGL, HSL or β-galactosidase (LacZ) using substrates containing radiolabeled triolein (TO), cholesteryl-oleate (CO), retinyl-palmitate, (RP), or phosphatidylcholine (PC) (*17*). (**E**) TG-hydrolase assay of cytosolic extracts of COS-7 cells expressing human ATGL (TTS-2.2), or β-galactosidase (LacZ) using a radiolabeled triolein substrate. Inset: The fluorescent lipase inhibitor NBD-HEHP binds to human ATGL (TTS-2.2) expressed in transfected COS-7 cells. Data are presented as means ± SD and represent at least three independent experiments. (***$P < 0.001$)

virus-40–transformed monkey kidney cells (COS-7) with cDNA clones expressing either murine histidine (His)–tagged ATGL or murine His-tagged HSL. Both enzymes were detected in the cytosolic supernatant and the membrane pellet fraction of transfected COS-7 cells by Western blotting (Fig. 1C). When extracts from transfected cells were preincubated with a fluorescent lipase inhibitor (NBD-HEHP) (*16*) and subsequently subjected to SDS–polyacrylamide gel electrophoresis (SDS-PAGE) analysis and fluorography (*17*), fluorescent signals were observed in positions corresponding to the expected molecular mass of ATGL (54 kD) and HSL (84 kD), (Fig. 1C). This fluorescent probe only reacts with enzymatically active Ser-lipases (*16*) which indicates that ATGL is enzymatically active in transfected COS-7 cells. To confirm this, we performed activity assays with radiolabeled lipid substrates (Fig. 1D) (*17*). In accordance with previous data (*18*, *19*), the cytosolic fractions of HSL-transfected cells exhibited increased TG-hydrolase (4.2-fold), cholesteryl ester hydrolase (23-fold), and retinyl ester hydrolase (2.3-fold) activities compared with LacZ-transfected cells. In contrast, cytosolic fractions of ATGL- transfected COS-7 cells hydrolized only the TG substrate (3.7-fold increase). Thus, ATGL is a TG-hydrolase, but in contrast to HSL, it does not hydrolyze cholesteryl or retinyl ester bonds. Similarly to murine ATGL, human ATGL (TTS-2.2) also exhibited marked lipolytic activity against a radiolabeled TG substrate and bound the lipase inhibitor NBD-HEHP (Fig. 1E).

Many TG lipases can hydrolyze more than one ester bond within the TG molecule, which results in the formation of monoglycerides or glycerol. For example, HSL cleaves TGs and DGs; however, its specific activity for DGs is 10 times that for TGs (*20*). In contrast, ATGL exhibited only very weak activity against a radiolabeled DG substrate compared with the TG substrate (Fig. 2A). Low DG-hydrolase activity of ATGL was also confirmed in experiments in which we measured the relative abundance of lipolytic reaction products (Fig. 2B) (*17*). Compared with control extracts of LacZ-transfected COS-7 cells, extracts from ATGL and HSL-transfected cells showed higher acyl-hydrolase activities (FA release) by factors of 7.5 and 10, respectively, when a [9,10-^{3}H(*N*)]-labeled triolein–substrate was used. In the presence of ATGL, the accumulation of DGs was increased 21-fold, which suggests that the enzyme predominantly hydrolyzed the first ester bond of TGs. In contrast, no DG accumulation was observed in lipolysis assays with cytosolic extracts from HSL-transfected cells. Monoglyceride levels were slightly increased in both ATGL- and HSL-transfected cells. From the molar ratios of DG and MG accumulation versus FA release we calculated that ~90% of the FA molecules released by the action of ATGL originate from the hydrolysis of TGs in the first ester bond. In contrast, in the presence of HSL, most FA originated from all three ester bonds resulting in glycerol formation. Thus, ATGL and HSL have different substrate-specificities within the lipolytic cascade, which suggests that they might act coordinately in the catabolism of TGs.

This hypothesis was confirmed by the product profiles generated in triolein hydrolysis assays using combined extracts of LacZ-, ATGL-, or HSL-transfected cells. Relative to extracts from LacZ-transfected cells, the acyl-hydrolase activity was increased in equal volume mixtures of HSL/LacZ extracts (4.8-fold), ATGL/LacZ extracts (4-fold), and ATGL/HSL extracts (16-fold) (Fig. 2C). The accumulation of DGs was increased 12.5-fold when LacZ/ATGL extracts were used and reduced to basal levels with ATGL/HSL extracts. We speculate that during the lipolytic breakdown of TGs, ATGL is predominantly responsible for the initial step of TG hydrolysis and provides DG substrate for the subsequent action of HSL, namely, the conversion of DGs into monoglycerides. In support of this model, the total acyl-hydrolase activity (FA release) in extracts containing ATGL/HSL was nearly 2 times the sum of the individual activities (Fig. 2C). During the final step of lipolysis, monoglycerides are converted to FA and glycerol by monoglyceride lipase (*21*).

The expected intracellular localization of a lipase involved in TG mobilization would be on lipid droplets. To determine whether this is true for ATGL, we constructed an adenovirus vector encoding His-tagged mouse ATGL and used it to infect 3T3-L1 adipocytes at day 8 of differentiation. Western blotting analysis revealed that the majority of ATGL protein (~50%) was present in the cytoplasm (Fig. 3A). However, a distinct fraction of ATGL (~10%) was found

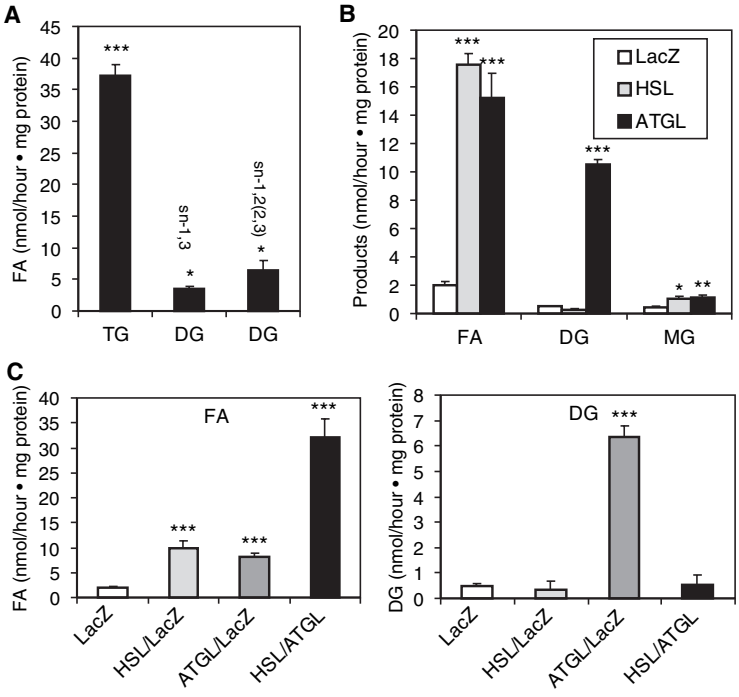


Fig. 2. Role of ATGL within the TG hydrolysis cascade. (**A**) Cytosolic extracts of HepG2 cells infected with an adenovirus construct expressing murine His-tagged ATGL (ATGL-Ad), or LacZ (LacZ-Ad) were incubated with radiolabeled TG or DG substrates. ATGL-mediated acyl-hydrolase activity was normalized by the activity measured in LacZ-Ad–infected cells. (**B**) Accumulation of reaction products during ATGL and HSL-mediated lipolysis. Cytosolic extracts of COS-7 cells transiently transfected with His-tagged LacZ, ATGL, or HSL were incubated with radiolabeled triolein. Lipids were extracted and separated by TLC, and the accumulation of free fatty acids (FA), digylcerides (DG), and monoglycerides (MG) was determined by liquid scintillation counting (*17*). (**C**) Effect of combined fractions of ATGL and/or HSL on acyl-hydrolase activity (FA) and diglyceride (DG) accumulation. Cytosolic extracts of COS-7 cells expressing LacZ were mixed 1:1 with extracts from cells expressing murine ATGL or HSL (ATGL/LacZ and HSL/LacZ) and compared with extracts prepared from a mixture of ATGL and HSL expressing cells (ATGL/HSL). Data are presented as means ± SD and represent three independent experiments. (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$)

tightly associated with lipid droplets of adipocytes even after extensive purification of the droplets. The amount of lipid droplet-associated ATGL was not affected by the stimulation of lipolysis with isoproterenol. In addition, fluorescence microscopy revealed that a green fluorescent protein-ATGL fusion protein localizes to the lipid droplet when expressed in 3T3-L1 adipocytes (fig. S3).

Further proof for a functional role of ATGL as a TG-hydrolase was provided by the fact that adenovirus-infected 3T3-L1 cells expressing ATGL released higher levels of FA (5-fold) and glycerol (1.8-fold) compared with LacZ-infected cells under basal conditions (Fig. 3B). After isoproterenol stimulation, FA release was increased 1.8-fold and glycerol release 2.9-fold. Additionally, ATGL-overexpression caused an increase in the cellular steady-state levels of DGs (fig. S4). Thus, ATGL in adipocytes can markedly augment both basal and isoproterenol-stimulated lipolysis. In contrast, silencing ATGL gene expression by siRNA (Fig. 3C) or antisense-RNA (fig. S5) markedly decreased the release of FA and glycerol from stimulated and nonstimulated 3T3-L1 adipocytes.

We next tested the effect of a rabbit polyclonal antibody against mouse ATGL (ATGL-IgG) on the enzyme activity of adipose tissue extracts from wild-type and HSL-deficient mice. In comparison with rabbit nonimmune IgG (NI-IgG), ATGL-IgG inhibited the cytosolic acyl-hydrolase activity in white and brown fat of wild-type mice by 64% and 71%, respectively (Fig. 3D). In white and brown adipose tissue of HSL-deficient mice, the activity was decreased by 75% and 74%, respectively. Thus the combined deficiency of HSL and ATGL in adipose tissue causes a loss of more than 90% of the acyl-hydrolase activity observed in wild-type adipose tissue. Compared with wild-type adipose tissue, the ATGL-mediated acyl-hydrolase activity and ATGL mRNA levels were not up-regulated in HSL-deficient adipose tissue (*22*).

Considering the central role of PKA in the regulation of fat cell lipolysis we tested whether ATGL is a target for PKA-mediated phosphorylation. As described in SOM (fig. S6), ATGL can be phosphorylated, but in contrast to HSL, this modification is not mediated by PKA.

In summary, our findings suggest that ATGL is an important component of the lipolytic process and the mobilization of lipid stores in mammals. It is responsible for the initial step in TG catabolism. Accordingly, the inhibition of ATGL offers a potential therapeutic approach to control FA release from adipose tissue in patients with insulin resistance.

Note added in proof: Two manuscripts in press present findings that the hormonally and nutritionally regulated protein desnutrin (*23*) and the TG-hydrolace inducible phospholipase-A2-ζ (*24*) are identical to ATGL.

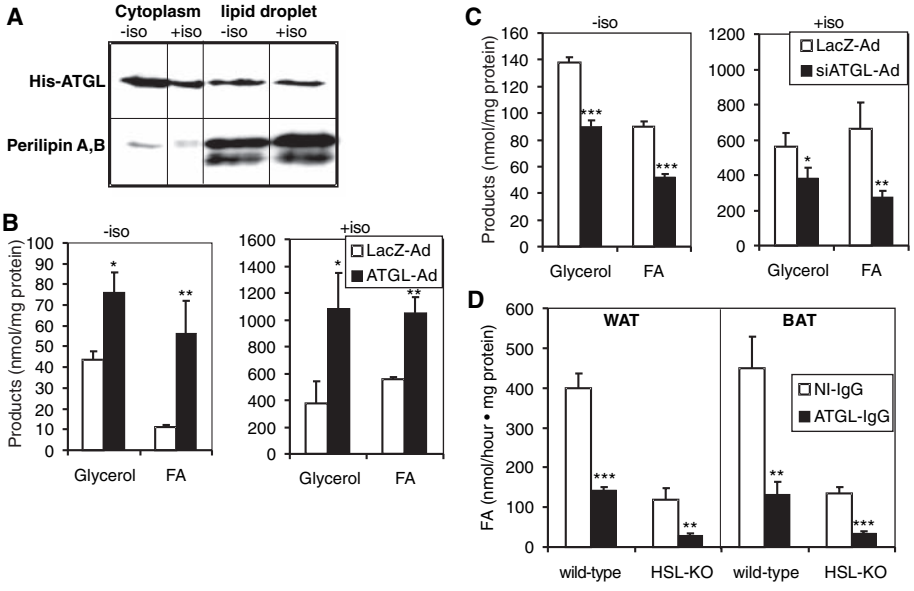


Fig. 3. Cellular localization and physiological function of ATGL in 3T3-L1 adipocytes. Recombinant adenovirus coding for LacZ (LacZ-Ad), His-tagged murine ATGL (ATGL-Ad) or a short interferring RNA for murine ATGL (siATGL-Ad) were used to infect 3T3-L1 adipocytes on day 8 of differentiation, and experiments were performed 2 days after infection (*17*). Before harvesting, cells were incubated in Dulbecco's minimum essential medium (DMEM)/2% FA-free bovine serum albumin (BSA) in the absence or presence of 10 μM isoproterenol (–/+ iso) for 1 hour (A) or 2 hours (B and C). (**A**) ATGL-protein was detected in the cytoplasmic fraction (10 μg of total protein) and in isolated lipid droplets (2 μg of total protein) of ATGL-Ad–infected 3T3-L1 adipocytes with an anti-His monoclonal antibody. Purification of lipid droplets was monitored by the enrichment of perilipin (>70-fold) using a rabbit polyclonal antibody against perilipin A and B. Release of glycerol and FA into the culture medium of 3T3-L1 adipocytes infected with ATGL-Ad (**B**) or siATGL-Ad (**C**). LacZ-Ad–infected cells were used as a control. Data are presented as means ± SD and represent three independent experiments. (**D**) Inhibition of cytosolic acyl-hydrolase activity in WAT and BAT by a polyclonal antibody against mouse ATGL (ATGL-IgG) measured by using radiolabeled triolein as substrate. The activity in cytosolic extracts of adipose tissue from wild-type and HSL-KO mice was determined in the presence of rabbit nonimmune IgG (NI-IgG) or ATGL-IgG. Data are presented as means ± SD and represent two independent experiments. WAT and BAT were obtained from three HSL-KO and three wild-type mice in each experiment. (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$)

References and Notes

1. R. N. Bergman *et al.*, *J. Invest. Med.* **49**, 119 (2001).
2. E. E. Blaak, *Proc. Nutr. Soc.* **62**, 753 (2003).
3. G. Boden, G. I. Shulman, *Eur. J. Clin. Invest.* **32**, (suppl. 3), 14 (2002).
4. P. Arner, *Diabetes Metab. Res. Rev.* **18** (suppl. 2), S5 (2002).
5. S. Collins, R. S. Surwit, *Recent Prog. Horm. Res.* **56**, 309 (2001).
6. C. Sztalryd *et al.*, *J. Cell Biol.* **161**, 1093 (2003).
7. S. P. Wang *et al.*, *Obes. Res.* **9**, 119 (2001).
8. R. Zimmermann *et al.*, *J. Lipid Res.* **44**, 2089 (2003).
9. H. Okazaki *et al.*, *Diabetes* **51**, 3368 (2002).
10. G. Haemmerle *et al.*, *J. Biol. Chem.* **277**, 4806 (2002).
11. R. L. Tatusov *et al.*, *Nucleic Acids Res.* **29**, 22 (2001).
12. A. Bateman *et al.*, *Nucleic Acids Res.* **32**, D138 (2004).
13. P. R. Shewry, *Ann. Bot. (London)* **91**, 755 (2003).
14. K. Athenstaedt, G. Daum, *J. Biol. Chem.* **278**, 23317 (2003).
15. A. Dessen *et al.*, *Cell* **97**, 349 (1999).
16. O. V. Oskolkova, R. Saf, E. Zenzmaier, A. Hermetter, *Chem. Phys. Lipids* **125**, 103 (2003).
17. Materials and methods are available on *Science* Online.
18. S. J. Yeaman, G. M. Smith, C. A. Jepson, S. L. Wood, N. Emmison, *Adv. Enzyme Regul.* **34**, 355 (1994).
19. S. Wei *et al.*, *J. Biol. Chem.* **272**, 14159 (1997).
20. G. Fredrikson, P. Stralfors, N. O. Nilsson, P. Belfrage, *J. Biol. Chem.* **256**, 6311 (1981).
21. G. Fredrikson, H. Tornqvist, P. Belfrage, *Biochim. Biophys. Acta* **876**, 288 (1986).
22. R. Zimmermann, R. Zechner, unpublished data.
23. J. A. Villena, S. Roy, E. Sarkadi-Nagy, K.-H. Kim, H. S. Sul, *J. Biol. Chem.* **279**, 47066 (2004); 10.1074/jbc.M403855200.
24. C. M. Jenkins *et al.*, *J. Biol. Chem.*, 10.1074/jbc.M407841200.

25. This work was supported by the Austrian Federal Ministery of Education, Science, and Culture (G.O.L.D., Genomics of Lipid-Associated Disorders and B.I.N., Bioinformatics Network) and by the Austrian Fonds zur Förderung der Wissenschlaftlichen Forschung (SFB Biomembranes F00701 and F007013). The authors thank G. Hoefler and M. Asslaber for the provision of human tissue biopsies, R. Schreiber and S. Eder for technical assistance, and E. Zechner for critically reviewing the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/306/5700/1383/DC1
Materials and Methods
Figs. S1 to S6
References and Notes

26 May 2004; accepted 17 September 2004

Nuclear Pore Complex Structure and Dynamics Revealed by Cryoelectron Tomography

Martin Beck, Friedrich Förster, Mary Ecke, Jürgen M. Plitzko, Frauke Melchior,* Günther Gerisch, Wolfgang Baumeister,† Ohad Medalia†

Nuclear pore complexes (NPCs) are gateways for nucleocytoplasmic exchange. To analyze their structure in a close-to-life state, we studied transport-active, intact nuclei from *Dictyostelium discoideum* by means of cryoelectron tomography. Subvolumes of the tomograms containing individual NPCs were extracted in silico and subjected to three-dimensional classification and averaging, whereby distinct structural states were observed. The central plug/transporter (CP/T) was variable in volume and could occupy different positions along the nucleocytoplasmic axis, which supports the notion that it essentially represents cargo in transit. Changes in the position of the CP/T were accompanied by structural rearrangements in the NPC scaffold.

Nuclear pore complexes (NPCs) mediate the exchange of macromolecules between the nucleus and the cytoplasm. These large assemblies (~120 megadaltons in metazoa) are constructed from about 30 different proteins, the nucleoporins (*1*). Functional and structural characterization of NPCs is challenging, due mainly to their sheer size. Structures of isolated NPCs have been resolved to 12 nm by means of cryo-EM (*2*), albeit with a nonisotropic resolution (see below). The structural analysis of active NPCs has obvious advantages over studies with isolated and detergent-extracted NPCs: The procedures that are traditionally used for the purification of NPCs are susceptible to loss of transport machinery components and cargo. In principle, cryoelectron tomography (cryo-ET) allows one to analyze the three-dimensional (3D) architecture of organelles or even whole cells embedded in vitreous ice, i.e., in a close-to-life state (*3*, *4*). Therefore, we have applied cryo-ET to whole *Dictyostelium discoideum* nuclei, which are relatively small (~2 μm) and can be isolated by a gentle procedure. In a projection image of such a nucleus after vitrification, an intact nuclear envelope is evident (Fig. 1A). These nuclei were fully competent for active nuclear import (Fig. 1, B and C; fig. S1), in a manner similar to permeabilized cells (*5*, *6*).

We acquired 16 tilt series of frozen-hydrated *Dictyostelium* nuclei and reconstructed the respective volumes. The framed area of the nucleus shown in Fig. 1A was used for recording a tomogram. Three different *x-y* slices (along the *z* axis) through this tomogram are shown (Fig. 1D). Both the outer nuclear membrane and a patch of connected endoplasmic reticulum are decorated with globular complexes that resemble 80*S* ribosomes in size and shape (~27 nm). In slices that are approximately perpendicular to the nuclear envelope, but more obviously in grazing slices, individual NPCs are clearly discernible. After surface rendering of the tomograms (Fig. 1E), clear pictures of individual NPCs were obtained. The canonical features such as cytoplasmic filaments, the three rings (cytoplasmic, lumenal spoke, and nuclear rings, respectively), parts of the basket, and the central plug/transporter (CP/T) were visible without any postprocessing.

We extracted 267 NPC-containing subvolumes from our tomograms. Given that our NPCs are transport-competent, one would expect them to be arrested in a variety of different transport states. Averaging without prior classification would therefore be expected to emphasize nonvariable features, whereas variable features would be deemphasized or even eliminated. Averaging procedures resulted in an isotropically sampled 3D density map of the NPC with a resolution of 8 to 9 nm. The cytoplasmic face (Fig. 2A)

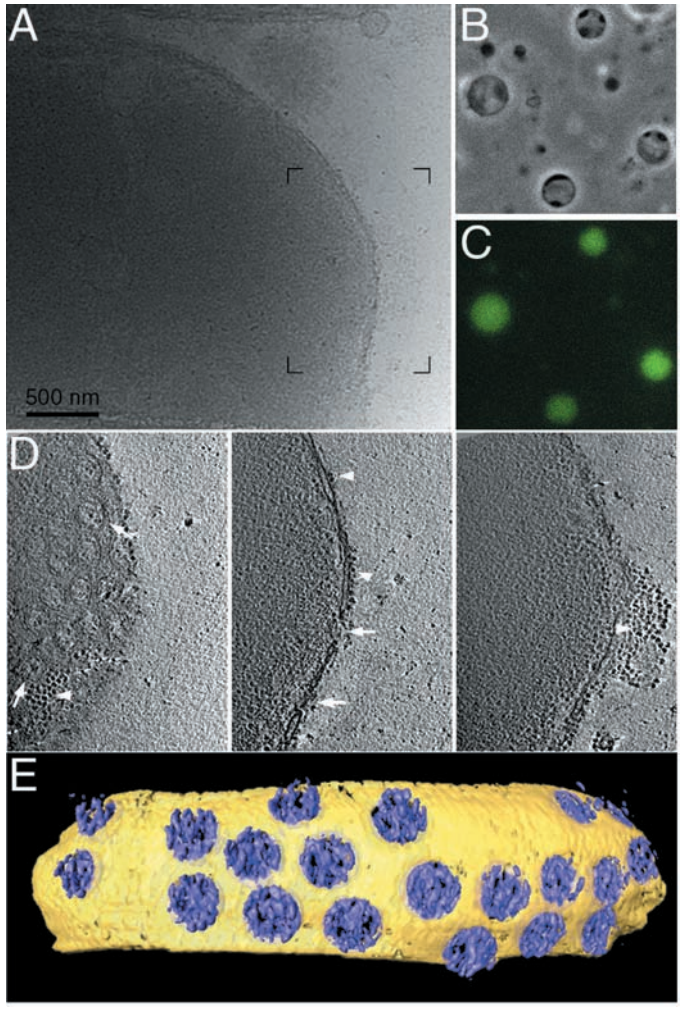


Fig. 1. Cryo-ET of transport-competent nuclei. (**A**) Transmission electron micrograph of a vitrified *Dictyostelium* nucleus. The image was recorded after acquisition of a complete tilt series; the frame marks the area representative for the reconstruction shown in (D). (**B**) Phase-contrast image and (**C**) the corresponding fluorescence image showing uptake of the transport substrate (FITC-BSA-NLS) into isolated, enriched nuclei. (**D**) Three-dimensional reconstruction of an intact nucleus. Three sequential *x-y* slices of 10 nm thickness along the *z* axis through a typical tomogram are indicated. Different orientations of NPCs are shown: top-views (left) and side-views (right, arrows). Ribosomes connected to the outer nuclear membrane are visible, as is a patch of rough ER (right, arrowheads). (**E**) Surface-rendered representation of a segment of nuclear envelope (NPCs in blue, membranes in yellow). The dimensions of the rendered volume are 1680 nm × 984 nm × 558 nm. The number of NPCs was ~45/μm².

Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany.

*Present address: Department of Biochemistry, University of Göttingen, D-37073 Göttingen, Germany.
†To whom correspondence should be addressed. E-mail: baumeist@biochem.mpg.de (W.B.) and omedalia@biochem.mpg.de (O.M.)

shows eight cytoplasmic filaments (~35 nm) arranged around the central channel. They protrude from the cytoplasmic ring and point toward the center of the structure. In spite of their flexible nature, these filaments show a distinct shape with a pointed kink. The CP/T comprises two overlapping, sphere-shaped densities. The smaller one (~20 nm) is in-plane with the cytoplasmic filaments; the larger one (~40 nm) is located farther down within the central channel (Fig. 2A). The most conspicuous feature on the nuclear face (Fig. 2B) is the nuclear basket. The nuclear filaments that connect the distal ring to the nuclear ring do not appear to be entirely straight but rather appear bent in the proximity of the nuclear ring (Fig. 2C, nuclear basket). The cytoplasmic as well as the nuclear filaments appear to be more delicate than in previous work that used metal coating, a technique that gives more prominence to filiform structures (*7*). The dimensions of the main features are revealed in a cutaway view without the CP/T (Fig. 2C). If one considers the corresponding dimensions that are known from different organisms, the diameter of the NPC from *Dictyostelium* is more similar to that of *Metazoa* than to yeasts (*7*, *8*), whereas the stack of rings is less elongated in the direction of the nucleocytoplasmic axis when compared with structures obtained previously by cryo-EM (*2*, *9*, *10*). The nuclear envelopes or detergent-extracted NPCs that were investigated in these studies assume a preferred orientation on the EM-grid. Consequently, the resulting structures were not isotropically sampled and lack information (the "missing cone") that leads to an artificial elongation along the *z* axis (*11*). In contrast, the NPCs of an intact nucleus assume free spatial orientations that represent all Eularian angles (fig. S3B). Gold-labeling experiments have shown that most of the components of the NPC (at least 18 different nucleoporins) are localized symmetrically in relation to an imaginary central plane through the lumenal spoke ring (*12*). In fact, when one looks at a slice oriented parallel to the nucleocytoplasmic axis, the upper and lower part of the lumenal spoke ring appear similar, whereas the CP/T appears rather asymmetrical (fig. S2B, right).

The CP/T located within the central channel has been described previously (*10*, *13*), but its role is still a matter of speculation. It is not found in every NPC (*13*), possibly because it is lost during the isolation procedure. Here, we found the CP/T in almost all NPCs that were examined, but its size, shape, and position varied substantially (fig. S2). Consequently, the average is rather featureless. We performed a quantitative analysis of this substructure in individual NPCs. The CP/Ts were extracted in silico and the occupied volumes, as well as the positions of the centers of gravity, were calculated. The analysis of the occupied volumes (Fig. 3A) supports the notion that the CP/T makes up, at least in part, cargo complexes arrested during translocation (*2*). The distribution of the centers of gravity along the nucleocytoplasmic axis indicates that two preferred positions of mass within the central channel exist, which are likely to correspond to different NPC states (Fig. 3B). It provides a means for an objective classification of individual NPCs. In order to visualize differences in the two NPC states,

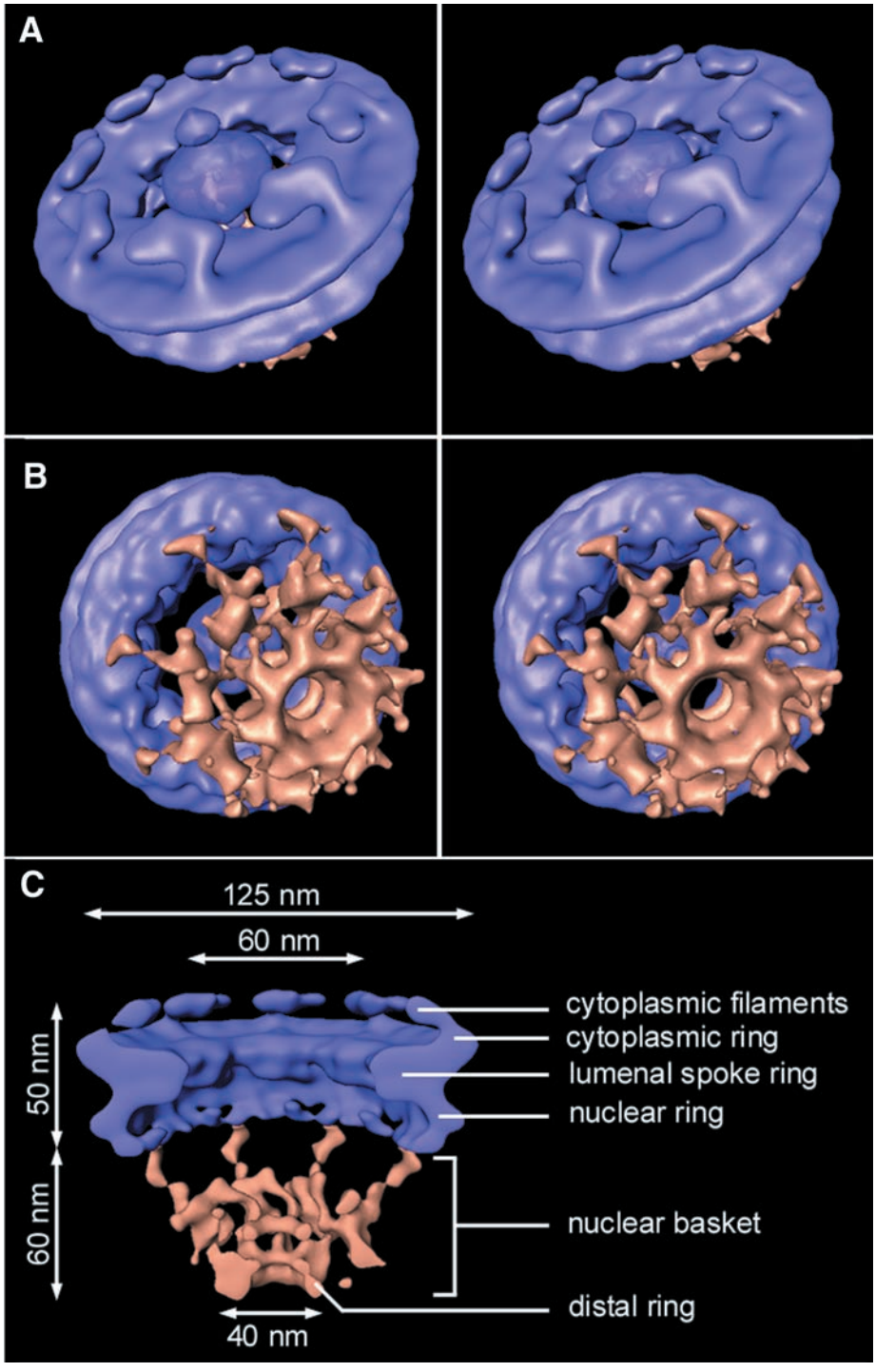


Fig. 2. Structure of the *Dictyostelium* NPC. (**A**). Cytoplasmic face of the NPC in stereo view. The cytoplasmic filaments are arranged around the central channel; they are kinked and point toward the CP/T. (**B**) Nuclear face of the NPC in stereo view. The distal ring of the basket is connected to the nuclear ring by the nuclear filaments. (**C**) Cutaway view of the NPC with the CP/T removed. The dimensions of the main features are indicated. All views are surface-rendered (nuclear basket in brown).

we divided our stack of particles into two classes: one designated a "cytoplasmic filament class" (CF class), and the second, a "lumenal spoke ring class" (LR class). Probability clouds for both classes have been calculated from the positions of the centers of gravity (Fig. 3C). The variance for the CF class is significantly smaller than for the LR class, which suggests that the position of the mass surrounded by the lumenal spoke ring is more diffuse. The two stacks of NPCs assigned to the CF class or the LR class were averaged separately. Slices through the structures of CF- and LR classes are shown in Fig. 3, D and E. The slices oriented along the nucleocytoplasmic axis (left) show the different positions of the CP/T. The cytoplasmic filaments are well defined in the CF class and have an apparent length of ~35 nm. An elongated density connects them to the CP/T (arrowhead). As revealed by the surface-rendered view shown in Fig. 3F (left), they are kinked, with the kink pointing toward the CP/T, whereas their tips point to adjacent filaments. The cytoplasmic face of the LR class shows only the base of the cytoplasmic filaments (Fig. 3G, left). The densities of the cytoplasmic and lumenal spoke ring are similar for both classes (Fig. 3, D and E, as CR, LR). The eight spokes of the lumenal spoke ring point toward the central channel in the LR class, but they appear bent in the CF class. Moreover, structural changes of the nuclear ring, which is connected to the nuclear basket, can be seen (Fig. 3, D and E, NR). Substantial differences in the structure of the nuclear basket are revealed in the surface-rendered views in Fig. 3, F and G (right). In the CF class, the distal ring has an opening in the middle and a smaller diameter; however, in the LR class it is more massive, indicative of the presence of additional mass bound to NPCs of this class.

The major differences between CF and LR classes are shown in Fig. 4. The schematic illustration shows the cytoplasmic filaments connected to the CP/T in the CF class (Fig. 4A), as indicated by the thin connection evident in the slice in Fig. 3D (arrowhead). We suggest that in this class, some of the cytoplasmic filaments are engaged in interactions with cargo that is represented by the CP/T and have a preferred orientation. As a result, the structure of the CF class represents the shape of a filament that interacts with cargo. Since it is not likely that all of the cytoplasmic filaments interact at the same time, their density is slightly "diluted" when they are averaged with imposed 8-fold symmetry. In the LR class, the cargo is not situated in the plane with the cytoplasmic filaments, which faded almost entirely during averaging. This suggests that the filaments

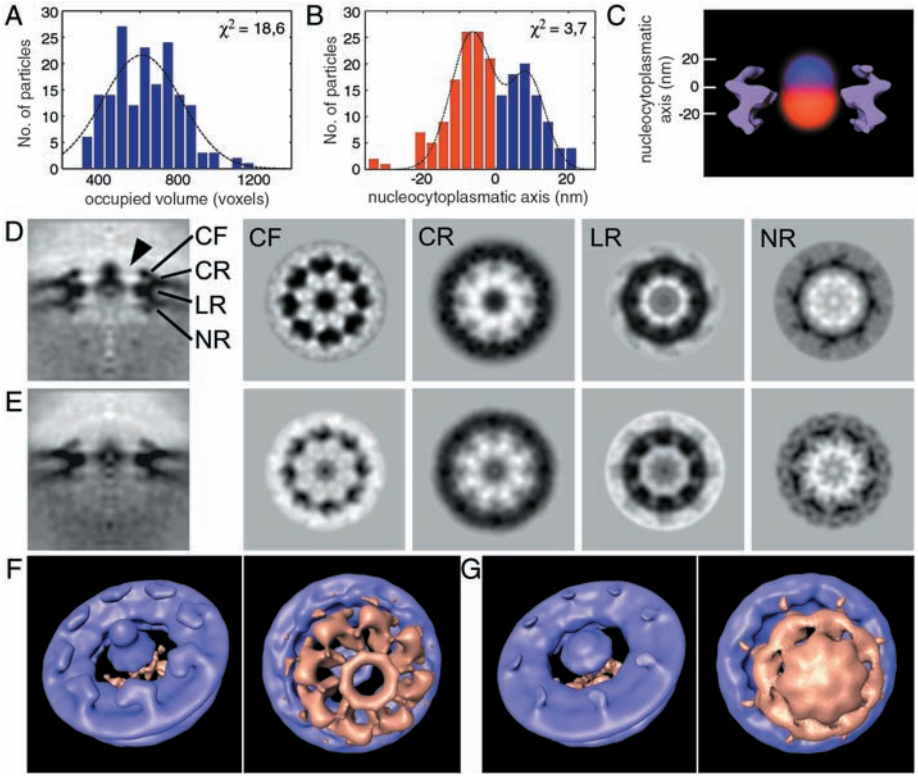


Fig. 3. Position of the CP/T correlates with structural changes of the NPC. (**A**) The distribution of the occupied volume in individual CP/Ts is significantly broader than the fitted Gaussian function, which would indicate that the CP/T in different NPCs has essentially the same components. The observed mass is thus likely to represent an average of different structures, rather than a constitutive substructure within the central channel of the NPC. (**B**) The distribution of the centers of gravity of individual CP/Ts along the nucleocytoplasmic axis shows two distinct peaks (LR class shaded in red). It matches a double-Gaussian function, which is indicative of two preferred positions of mass along the same axis. The quality of the fits in (A) and (B) was evaluated with a chi-square test. The disjunction criterion was chosen at the minimum between the two peaks and is also indicated in the average viewed in fig. S2B (right). (**C**) Probability clouds of the centers of gravity corresponding to both classes. The position of the centers of gravity is shown superimposed on a centered slice through the structure (CF class in blue, LR class in red). The center of gravity is better defined in the CF class. (**D**) Slices through the structure of the CF class. Centered slices along the nucleocytoplasmic axis (left) and slices in-plane with the cytoplasmic filaments (CF), cytoplasmic ring (CR), lumenal spoke ring (LR), and nuclear ring (NR), corresponding to 7 nm in thickness. The cytoplasmic filaments are connected to the CP/T by an elongated density (arrowhead). (**E**) Same as (D) but for the LR class. (**F**) Surface-rendered views of the CF class. Cytoplasmic face view (left) and nuclear face view (right). (**G**) Same as (F) but for the LR class (nuclear basket in brown).

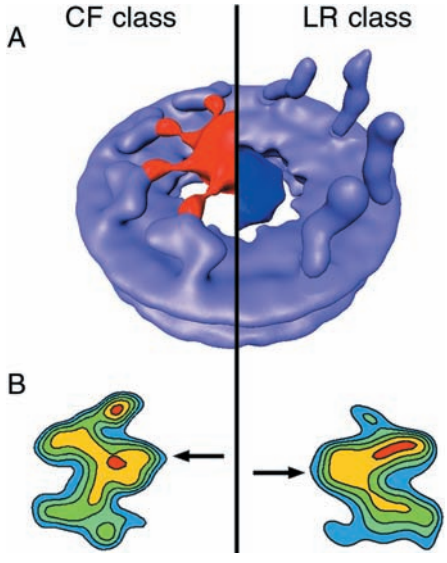


Fig. 4. Comparison of the two structural states of the NPC. (**A**) Schematic illustration of the structural changes of the cytoplasmic filaments. Surface-rendered views of the structure superimposed with the CP/Ts at a lower threshold (CF class in red, LR class in blue). In the CF class, the cytoplasmic filaments are in a defined orientation and interact with the CP/T; the latter is situated within the same plane. In the LR class, the CP/T is located in the plane with the lumenal spoke ring, and the disengaged cytoplasmic filaments are variable in shape and fade out in the average. Therefore, we added four with arbitrary shapes for the sake of completeness. (**B**) Contour-line-view of slices along the nucleocytoplasmic axis. The position of the narrowest constriction in the central channel is indicated for both classes (arrows).

that do not interact with the CP/T at this stage are variable in shape, while binding to cargo restricts their freedom of movement (*14*). The contour line view reveals that the narrowest constriction of the central channel is situated at the cytoplasmic side of the lumenal spoke ring in the CF class, even though it is at the nuclear side of the same ring in the LR class (Fig. 4B). These observations indicate that major rearrangements in the spokes might play a critical role in the translocation of cargo.

Both classes represent major structural states of the NPC. Since the CP/T is better defined in the CF class, this state might represent the slow incorporation or release of cargo complexes into or from the FXFG-framework residing in the central channel (*15*), which involves interaction with the cytoplasmic filaments. The more diffuse CP/T of the LR class indicates that cargo complexes can be found in various positions once they have entered the channel [in agreement with (*16*)]. Although an assignment of the classes to import, export or to predominant rate-limiting steps in both processes is not yet possible, the application of cryo-ET to transport-competent, intact nuclei holds great potential for a structural dissection of the key steps involved. The use of defined cargo and the trapping of distinct transport intermediates should ultimately enable us to arrive at a detailed mechanistic understanding of the nuclear pore complex.

References and Notes

1. M. Suntharalingam, S. R. Wente, *Dev. Cell* **4**, 775 (2003).
2. D. Stoffler *et al.*, *J. Mol. Biol.* **328**, 119 (2003).
3. W. Baumeister, *Curr. Opin. Struct. Biol.* **12**, 679 (2002).
4. O. Medalia *et al.*, *Science* **298**, 1209 (2002).
5. F. Melchior, B. Paschal, J. Evans, L. Gerace, *J. Cell Biol.* **123**, 1649 (1993).
6. Because the NPCs of spread *Xenopus* nuclear envelopes have a preferred orientation (*17*), an isotropically resolved structure cannot be obtained from such samples (although transport active), owing to the missing cone problem (*11*).
7. M. W. Goldberg, T. D. Allen, *J. Mol. Biol.* **257**, 848 (1996).
8. E. Kiseleva *et al.*, *J. Struct. Biol.* **145**, 272 (2004).
9. Q. Yang, M. P. Rout, C. W. Akey, *Mol. Cell* **1**, 223 (1998).
10. C. W. Akey, M. Radermacher, *J. Cell Biol.* **122**, 1 (1993).
11. K. Grunewald, O. Medalia, A. Gross, A. C. Steven, W. Baumeister, *Biophys. Chem.* **100**, 577 (2003).
12. M. P. Rout *et al.*, *J. Cell Biol.* **148**, 635 (2000).
13. E. Kiseleva, M. W. Goldberg, T. D. Allen, C. W. Akey, *J. Cell Sci.* **111**, 223 (1998).
14. N. Pante, U. Aebi, *Science* **273**, 1729 (1996).
15. B. Fahrenkrog, U. Aebi, *Nat. Rev. Mol. Cell Biol.* **4**, 757 (2003).
16. W. Yang, J. Gelles, S. M. Musser, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12887 (2004).
17. J. P. Siebrasse, R. Peters, *EMBO Rep.* **3**, 887 (2002).

18. We thank R. Hegerl for help with the image processing and A. Leis, V. Lucic, and P. Zwickl for critical reading of the manuscript.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1104808/DC1
Materials and Methods
Figs. S1 to S3
References and Notes

3 September 2004; accepted 24 September 2004
Published online 28 October 2004;
10.1126/science.1104808
Include this information when citing this paper.

Anabaena Sensory Rhodopsin: A Photochromic Color Sensor at 2.0 Å

Lutz Vogeley,[1] Oleg A. Sineshchekov,[3,5] Vishwa D. Trivedi,[3] Jun Sasaki,[3] John L. Spudich,[3,4]* Hartmut Luecke[1,2]*

Microbial sensory rhodopsins are a family of membrane-embedded photoreceptors in prokaryotic and eukaryotic organisms. Structures of archaeal rhodopsins, which function as light-driven ion pumps or photosensors, have been reported. We present the structure of a eubacterial rhodopsin, which differs from those of previously characterized archaeal rhodopsins in its chromophore and cytoplasmic-side portions. *Anabaena* sensory rhodopsin exhibits light-induced interconversion between stable 13-cis and all-trans states of the retinylidene protein. The ratio of its cis and trans chromophore forms depends on the wavelength of illumination, thus providing a mechanism for a single protein to signal the color of light, for example, to regulate color-sensitive processes such as chromatic adaptation in photosynthesis. Its cytoplasmic half channel, highly hydrophobic in the archaeal rhodopsins, contains numerous hydrophilic residues networked by water molecules, providing a connection from the photoactive site to the cytoplasmic surface believed to interact with the receptor's soluble 14-kilodalton transducer.

Over the past 4 years, microbial genomics has revealed a large family of photoactive, seven-transmembrane-helix retinylidene proteins called microbial rhodopsins in phylogenetically diverse species, including haloarchaea, proteobacteria, cyanobacteria, fungi, and algae (*1*–*4*). The first members of this family were discovered in halophilic archaea: the light-driven ion pumps bacteriorhodopsin and halorhodopsin and the phototaxis receptors sensory rhodopsins I and II. These four related haloarchaeal pigments are among the best-characterized membrane proteins in terms of structure and function, and nearly all of our knowledge of the properties of microbial rhodopsins, such as isomeric configuration and conformation of their chromophore, photochemical reactions, light-induced conformational changes in the protein, and function, derives from the study of these four, including atomic resolution structures that have been obtained for three of them (*5*–*9*). Studies of non-haloarchaeal rhodopsins, of which >800 are known to exist (*10*, *11*), are needed to examine the diversity of properties of this widespread family (*12*). *Anabaena* sensory rhodopsin, a recently discovered sensory representative outside of archaea (*2*), is well suited for exploration. It is the only bacterial sensory rhodopsin so far expressed in a photoactive form. Unlike the haloarchaeal sensory rhodopsins, which transmit signals to other integral membrane proteins, its function appears to involve modulation of a soluble cytoplasmic transducer, analogous to animal visual pigments (*2*).

In this study, we report the structure of the retinal-complexed protein at 2.0 Å resolution, obtained by X-ray diffraction of crystals grown in a cubic lipid phase (table S1). The overall membrane-embedded seven-helical structure is similar to those of the archaeal rhodopsins. However, distinct differences in the photoactive site prompted analysis of the isomeric configuration of the retinal and the photochemical reactions of the pigment.

Despite intense white-light illumination [light adaptation (*13*)] of the crystals before cryocooling and X-ray data collection, which results in a fully all-trans retinal configuration in bacteriorhodopsin, maps of the retinal and Schiff base region of *Anabaena* sensory rhodopsin show electron density incompati-

[1]Department of Molecular Biology and Biochemistry, [2]Department of Physiology and Biophysics and Department of Informatics and Computer Sciences, University of California, Irvine, CA 92697, USA. [3]Center for Membrane Biology, Department of Biochemistry and Molecular Biology, [4]Department of Microbiology and Molecular Genetics, University of Texas Medical School, Houston, TX 77030, USA. [5]Biology Department, Moscow State University, Moscow, Russia.

*To whom correspondence should be addressed. E-mail: hudel@uci.edu (H.L.) or john.l.spudich@uth.tmc.edu (J.L.S.)

ble with 100% all-trans retinal (Fig. 1A). Subsequent extractions and chemical structure determinations of retinal isomers from orange-illuminated (580-nm) and blue-illuminated (480-nm) *Anabaena* pigment showed light-induced shifts of the isomeric configuration. In the fully dark-adapted state, the all-trans form [absorption maximum (λ_{max}) of 549 nm in detergent-solubilized membranes] predominates [>75%, (Fig. 1B)]. Orange illumination rapidly shifted the pigment to a stable >80% 13-cis state (λ_{max} of 537 nm), and blue light rapidly increased the all-trans content toward the dark-adapted isomer ratio (Fig. 1B). Therefore, the relative amounts of *Anabaena* sensory rhodopsin with cis and trans chromophore configurations depend on the quality (color) of illumination and are shifted between the two forms by pulses of orange and blue illumination (Fig. 1C). This photochromic property provides a possible mechanism for single-pigment color sensing. Its two distinct groundstate species thermally interconvert with halftimes of ~100 min and ~300 min for the trans and cis forms, respectively; this is a fundamental difference from that of another color-sensitive microbial rhodopsin, the archaeal phototaxis receptor sensory rhodopsin I (*14*). Such relatively long-lasting color sensitivity is similar to that of the red/far-red photochromic states of phytochrome and may be used, in the *Anabaena* cell in analogy to phytochrome (*15*–*17*), to control expression of proteins required under either orange-light or blue-light illumination. The photochromic reactions are also similar to those between 11-cis and all-trans forms of invertebrate visual pigments, which have been suggested to reset the 11-cis state in a light-dependent manner (*18*).

Further detailed structural analysis of the active site revealed two alternate conformations for Lys^{210}. In one, its carbonyl oxygen forms a regular α-helical hydrogen bond with the peptide of Ser^{214}; in the other, its hydrogen bond donor is a nearby water (Wat^{502}) (Fig. 2B). Wat^{502} also connects helices B and G by bridging the hydroxyl of Ser^{214} with the backbone carbonyl of Ala^{40}. Multiple conformations of the residue-210 peptide may be facilitated by the presence of a π bulge at residues Ser^{209}, Lys^{210}, and Val^{211}, which is believed to soften the otherwise relatively rigid α helix (*5*, *19*). A further reduction of the α-helical character of this region stems from the replacement of the aspartic residue at position 206 (anionic Asp^{212}, which is part of the complex counterion in bacteriorhodopsin), highly conserved in archaeal rhodopsins, with a proline, Pro^{206} (Fig. 2A). Although the α helix on both sides of Pro^{206} is undisturbed by the loss of the peptide amide of the proline, the main-chain carbonyl of residue 202 accepts a hydrogen bond from the hydroxyl of reoriented Tyr^{51}. In other microbial rhodopsin structures, the Tyr^{51} hydroxyl forms a strong hydrogen bond with the anionic aspartate carboxyl. The rearrangement also results in a 1.3 Å movement of Wat^{402}, the water that bridges the protonated Schiff base and its counterion (*5*, *7*, *20*), toward the β-ionone ring of the retinal. Wat^{402} receives hydrogen bonds from the Schiff base (3.0 Å versus 2.6 Å in sensory rhodopsin II) and from the Trp^{76} indole while donating hydrogen bonds to the OD2 of Asp^{75} and, weakly, to the hydroxyl of Tyr^{51}. Further toward the extracellular side, the flexible guanidinium side chain of Arg^{72} points away from the Schiff base and toward the extracellular side, as in archaeal sensory rhodopsin II (*7*); however, here Arg^{72} is flanked by two histidines (His^{69} and His^{8}).

Comparison of the cytoplasmic half of *Anabaena* sensory rhodopsin with those of other microbial rhodopsins reveals markedly increased hydrophilicity in this region (Fig. 2B). The active site near the middle of the bilayer is connected to the cytoplasm via a hydrophilic path that contains at least four water molecules. A number of hydrophilic side chains interact with these water molecules to form an almost continuous hydrogen-bonded network from the Lys^{210} carbonyl to the cytoplasm over a distance of 19 Å: Lys^{210} – Wat^{502} – Ser^{214} – Asp^{217} of helix G; Ser^{86}, Thr^{90}, and Gln^{93} of helix C and the C-D loop; and Glu^{36} of helix B (Fig. 2B). In contrast, the cytoplasmic region of the haloarchaeal sensory rhodopsin II is entirely hydrophobic (*7*). Most notably, Phe^{86} in the archaeal protein occupies the space occupied by three water molecules and Ser^{86} in the center of the hydrophilic path of the *Anabaena* protein. This difference is consistent with the fundamentally different transducer interactions of

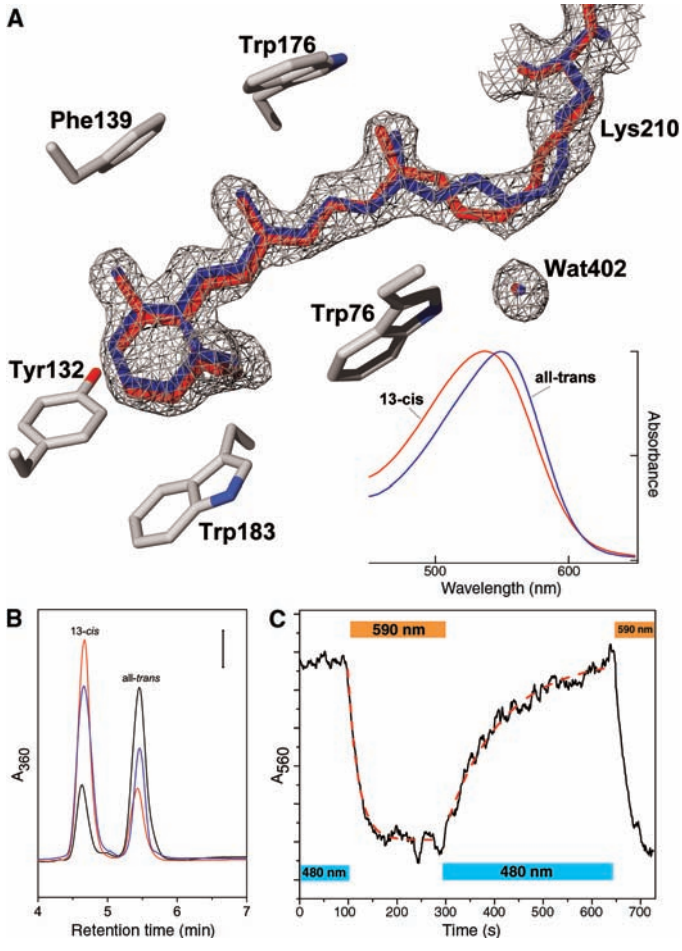


Fig. 1. The electron densities indicate a mixture of retinal isomers. (**A**) Annealed electron density omit map contoured at 1σ with 13-cis,15-syn (in red) and all-trans,15-anti (in blue) retinal models. The density suggests a mixture of all-trans,15-anti and 13-cis,15-syn retinal after white-light illumination (*13*). The conjugated π system of the retinylidene is more bent than in archaeal sensory rhodopsin II, with the distance from the Schiff base nitrogen to the β-ionone C1 reduced to 11.6 Å from 12.2 Å. The increased bent also causes an increase, by 0.6 Å, in the distance between the two tryptophan side chains (Trp^{76} and Trp^{176}) that sandwich the retinal in its binding site. (Inset) Absorption spectra of the 13-cis and trans forms of the pigment calculated from the measured spectra of the orange-illuminated and dark-adapted states using the isomer ratios depicted in (**B**). (**B**) Extraction of retinal isomers from orange-illuminated (580 ± 5 nm, 5 min, red line), blue-illuminated (480 ± 5 nm, 5 min, blue line), and dark-adapted (black line) pigments in detergent-solubilized (0.1% dodecylmaltoside) *Escherichia coli* membranes revealed a decrease of the fraction of 13-cis,15-syn retinal from 82% (orange) to 24% (dark). The units of the A_{360} axis (absorbance at 360 nm) are 2.0×10^{-3}, 1.2×10^{-3}, and 1.2×10^{-3} absorption units for the orange-illuminated, blue-illuminated, and dark-adapted samples, respectively. (**C**) Photoconversion between cis- and trans-forms under continuous monochromatic illumination. Absorbance at 560 nm, greater in the all-trans form compared to the 13-cis form, is used to monitor the wavelength-sensitive spectral transitions. The photoconversions follow approximately first-order kinetics as shown by the single exponential fits (dashed red curves) to the transitions during illumination through 10-nm band-pass interference filters centered at 590 nm or 480 nm, as indicated.

the *Anabaena* photoreceptor (soluble transducer) and haloarchaeal photoreceptor (membrane-embedded transducer) (*2*). For the latter, the cubic lipid phase crystal structure was used to predict the membrane-embedded surface of transducer interaction (*7*), later confirmed by the crystal structure of the receptor bound to a transducer fragment (*9*). The soluble *Anabaena* transducer (*2*) is thought to interact through the receptor's cytoplasmic surface. In the *Anabaena* photosensor, this surface is highly ordered, and all three loops that connect the transmembrane α helices (the A-B, C-D, and E-F loops) are structurally well defined, with conformations substantially different from those of bacteriorhodopsin and sensory rhodopsin II (Fig. 2C). Specifically, Gln^{93} is part of a four-residue insertion in the C-D loop relative to the archaeal receptor that results in an enlarged yet well-ordered cytoplasmic loop near the end of the hydrophilic path, a region likely to interact with the transducer.

As with most other membrane protein crystals prepared from the cubic lipid phase, long, tubular electron densities could be interpreted as lipid tails that form ordered, stacked bilayers in the crystal. Judging from the 13 lipid tails that could be built into electron density, it appears that, in contrast to earlier studies of cubic lipid phase crystals, this bilayer is not planar in *Anabaena* sensory rhodopsin crystals but rather undulates as a result of specific protein-protein interactions within and between bilayers (fig. S1).

The data shown here reveal two photochromic states of *Anabaena* sensory rhodopsin determined by the color of ambient light. The physiological function of the receptor is not yet known, but in cyanobacteria several physiological processes depend on light in the region of its absorption (*2*). For example, cyanobacteria adjust the pigment composition of their photosynthetic light-harvesting complexes based on the color of available light, a phenomenon called chromatic adaptation. Action spectra for chromatic adaptation show that orange light stimulates synthesis of phycocyanin, whereas shorter wavelength blue-green light activates synthesis of phycoerythrin (*21–23*). This color-sensitive pigment synthesis is generally assumed to be based on participation of two competitive receptor pigments with orange versus blue-green absorption maxima. However, the photochromic property of the *Anabaena* pigment shows that it is possible that such color sensing could be achieved by a single photoreceptor, namely the pigment in its two photo-interconvertible groundstates. The signaling mechanism could make use either of the ratio of the two stable groundstate forms or photochemical reaction of one of the forms, because in both cases the photointerconversion between the cis- and trans-forms of the pigment depends on the light quality.

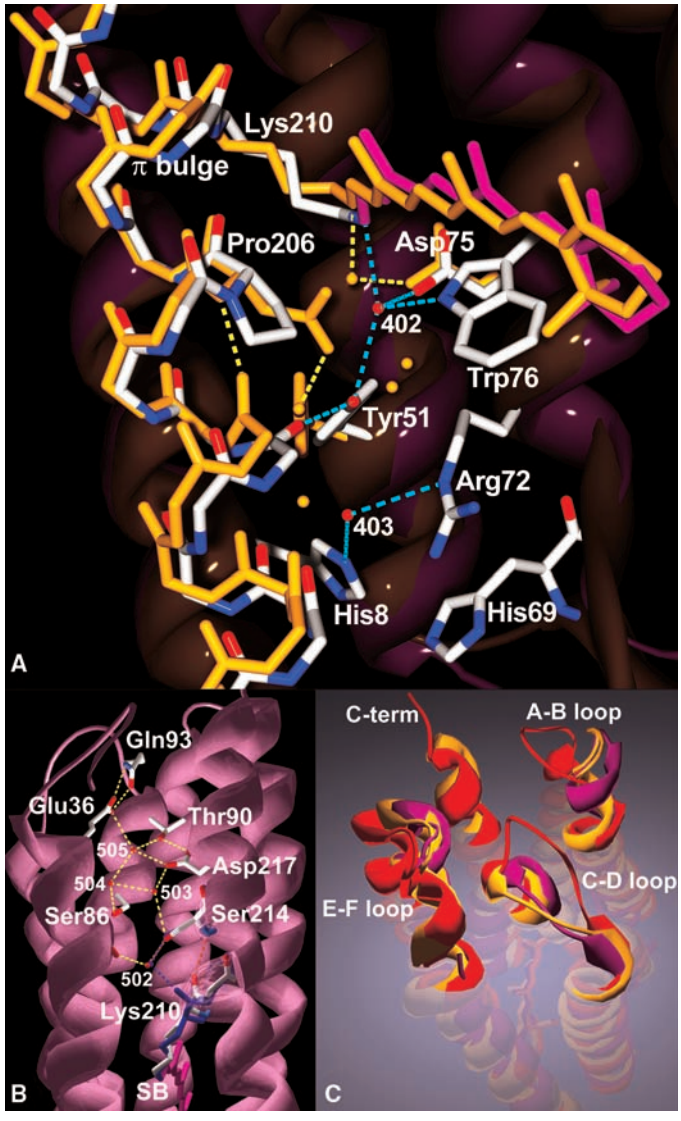


Fig. 2. Structural differences with archaeal rhodopsins. (**A**) The extracellular half of *Anabaena* sensory rhodopsin is shown as purple ribbon with CPK-colored atoms, magenta retinal near the top right, red water molecules, and turquoise hydrogen bonds, with residue numbering according to its sequence. The extracellular surface is near the bottom. The largest differences appear around the Asp-to-Pro mutation at position 206. For comparison, archaeal sensory rhodopsin II is shown in orange throughout with yellow hydrogen bonds. (**B**) The cytoplasmic half of the protein is markedly more hydrophilic than those of other microbial rhodopsins. The cytoplasmic surface is located at the top of the image and the retinal near the bottom. The peptide plane between residues 210 and 211 displays two alternate conformations. In one, Lys^{210} C=O accepts a regular intrahelical hydrogen bond from residue 214 N-H (Lys^{210} shown in blue; hydrogen bonds are shown as red dashed lines). In the other, it accepts a hydrogen bond from Wat^{502} (hydrogen bonds are shown as blue dashed lines). This alternate conformation results in an ~55° change in the orientation of the 210 peptide bond C=O vector, with a movement of the Lys^{210} carbonyl oxygen by 1.8 Å. Only the latter conformation completes a hydrogen bond chain that leads from Lys^{210} C=O at the active site via Wat^{502}, Ser^{214}, OH, and three more ordered waters (Wat^{503}, Wat^{504}, and Wat^{505}) held in place by the side chains of Asp^{217}, Ser^{86} (two alternate side-chain conformations, only one of which is shown for clarity), and Thr^{90} to the cytoplasmic surface near Glu^{36} of helix B and Gln^{93} in the C-D loop. (**C**) A comparison of the loop structures that define the respective cytoplasmic surfaces reveals large differences between the surfaces of archaeal sensory rhodopsin II (orange) and bacteriorhodopsin (purple) and the surface of *Anabaena* sensory rhodopsin (red), which is thought to interact with its soluble transducer. In particular, the A-B and C-D loops of the *Anabaena* protein are packed entirely differently, with relative backbone movements of 10 Å and 7 Å, respectively. The C-D loop contains surface-exposed Phe^{94}/Ile^{95}, Lys^{96}/Lys^{97}, and Trp^{99} side chains, and because of a four-residue insertion relative to sensory rhodopsin II, the loop protrudes 6 Å further into the cytoplasmic space.

References and Notes

1. These pigments, also known as type 1 rhodopsins, are found in each of the three domains of life (i.e., Archaea, Bacteria, and Eucarya), including haloarchaea, proteobacteria, cyanobacteria such as *Anabaena*, fungi, and algae (*2–4*, *24*).
2. K. H. Jung, V. D. Trivedi, J. L. Spudich, *Mol. Microbiol.* **47**, 1513 (2003).
3. O. Béjà *et al.*, *Science* **289**, 1902 (2000).
4. O. A. Sineshchekov, K. H. Jung, J. L. Spudich, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8689 (2002).
5. H. Luecke, B. Schobert, H. T. Richter, J. P. Cartailler, J. K. Lanyi, *J. Mol. Biol.* **291**, 899 (1999).
6. M. Kolbe, H. Besir, L. O. Essen, D. Oesterhelt, *Science* **288**, 1390 (2000).
7. H. Luecke, B. Schobert, J. K. Lanyi, E. N. Spudich, J. L. Spudich, *Science* **293**, 1499 (2001).
8. A. Royant *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10131 (2001).
9. V. I. Gordeliy *et al.*, *Nature* **419**, 484 (2002).
10. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
11. J. L. Spudich, K. H. Jung, in *Handbook of Photosensory Receptors*, W. Briggs, J. L. Spudich, Eds. (Wiley, Weinheim, Germany), in press.

12. O. A. Sineshchekov, J. L. Spudich, *Photochem. Photobiol. Sci.* **3**, 548 (2004).
13. The traditional notion of light and dark adaptation of microbial rhodopsins (LA and DA, illumination with intense white light and relaxation to thermodynamic equilibrium in the dark, respectively) stems from decades of research on bacteriorhodopsin. For wild-type bacteriorhodopsin, LA produces nearly 100% all-trans retinal, whereas DA yields ~40% all-trans and 60% 13-cis,15-syn retinal. Because for *Anabaena* sensory rhodopsin, neither LA nor DA may represent a physiologically relevant state, we prefer to refer to orange-illuminated or blue-illuminated chromophores. The former encompasses cases in which the incident light is within the pigment's absorption range but of substantially longer wavelength than the λ_{max} of the chromophore and the latter cases where the light is of substantially shorter wavelength.
14. Sensory rhodopsin I produces an attractant signal in response to orange-light–induced trans-to-cis isomerization, and its photocycle contains a transient 13-cis blue-shifted photointermediate. This intermediate's cis-to-trans photoreaction from near-ultraviolet (UV) light generates a repellent signal (*25*). Sensory rhodopsin I therefore detects the presence of near-UV light in an orange-light background over the few seconds' duration of its photocycle. *Anabaena* sensory rhodopsin, in contrast, exhibits two distinct dark groundstate spectral species, each of which is stable for several orders of magnitude longer than their flash-induced photocycles (*26*).
15. Phytochromes exhibit red-absorbing and far-red–absorbing forms that control a variety of phenomena in plants, such as flowering and circadian rhythms. As described for *Anabaena* sensory rhodopsin here, the two forms of phytochrome are each stable in the dark over long periods and are rapidly photointerconverted, properties that provide color-sensitive physiological responses (*16*, *17*, *27*).
16. H. Wang, X. W. Deng, in *The Arabidopsis Book*, C. R. Somerville, E. M. Meyerowitz, Eds. (American Society of Plant Biologists, Rockville, MD, 2002), pp. 1–28.
17. P. Gyula, E. Schäfer, F. Nagy, *Curr. Opin. Plant Biol.* **6**, 446 (2003).
18. W. Gärtner, in *Handbook of Biological Physics*, D. G. Stavenga, W. J. de Grip, E. N. Pugh Jr., Eds. (Elsevier, Amsterdam, 2000), vol. 3, pp. 297–388.
19. J. P. Cartailler, H. Luecke, *Structure* **12**, 133 (2004).
20. H. Luecke, H. T. Richter, J. K. Lanyi, *Science* **280**, 1934 (1998).
21. R. MacColl, *J. Struct. Biol.* **124**, 311 (1998).
22. A. R. Grossman, D. Bhaya, Q. He, *J. Biol. Chem.* **276**, 11449 (2001).
23. B. Singh, V. S. Chauhan, S. Singh, P. S. Bisen, *Curr. Microbiol.* **43**, 265 (2001).
24. J. L. Spudich, C. S. Yang, K. H. Jung, E. N. Spudich, *Annu. Rev. Cell Dev. Biol.* **16**, 365 (2000).
25. J. L. Spudich, R. A. Bogomolni, *Nature* **312**, 509 (1984).
26. O. A. Sineshchekov *et al.*, in preparation.
27. G. Wagner, in *ESP Review Series on Photobiology*, D.-P. Häder, M. Lebert, Eds. (Elsevier, Amsterdam, 2001), pp. 421–448.
28. Supported by NIH grant nos. R01-GM59970 (H.L.), R01-GM067808 (H.L.) and R37-GM27750 (J.L.S.); NSF grant no. 0091287 (J.L.S.); a Welch Investigator Award (J.L.S.); and a Bessel Award from the Alexander-von-Humboldt Foundation (H.L.). The atomic coordinates and structure factors of *Anabaena* sensory rhodopsin are available at the Protein Data Bank with code 1XIO.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1103943/DC1
Materials and Methods
Fig. S1
Table S1
References and Notes

11 August 2004; accepted 22 September 2004
Published online 30 September 2004;
10.1126/science.1103943
Include this information when citing this paper.

New Products

JEOL USA
For more information
978-535-5900
www.jeol.com
http://science.labvelocity.com

NMR Without Deuterated Solvents

An automated No-D (no-deuterium proton) nuclear magnetic resonance (NMR) spectrometry function provides a fast, easy way for the synthetic organic chemist to run NMR experiments without using deuterated solvents. The entire No-D NMR process is automated from setup to data processing with the advanced software of the ECA and ECX series NMR spectrometers. No-D NMR has been underused as a means of analyzing pure compounds and reaction mixtures due to the perception that it is difficult to obtain NMR data in the presence of non-deuterated solvents. However, researchers can now easily and selectively wipe out unwanted solvent peaks in protonated solvents by using suppression techniques common to all modern NMR spectrometers. The automated No-D NMR function streamlines high-quality data acquisition with simple push-button operation, eliminating the need for time-consuming manual adjustment of parameters and experimental setups. It automatically performs gradient shimming, solvent suppression, referencing, and temperature control.

Partek
For more information
636-498-2329
www.partek.com
http://science.labvelocity.com

Statistical Analysis and Data Visualization

Partek Pro version 6.0 is a statistical analysis and interactive data visualization system that is well-suited for analysis of experiments and studies involving high-throughput and high-content technologies. Designed with the scientist in mind, it provides powerful yet easy-to-use tools that are fast enough for today's high-dimensional data. It is available for Windows, IRIX, Solaris, and Linux systems.

IBA
For more information
+49-551-50672-114
www.iba-go.com
http://science.labvelocity.com

Magnet-Assisted Transfection

The MATra product line allows efficient transfection of cells in culture through the use of magnetic nanoparticles (MagTag). With this new, easy-to-handle technique, nucleic acids, such as plasmid DNA, oligonucleotides, or small interfering-RNA are first associated with magnetic particles. Through the exploitation of magnetic force, the full nucleic acid is then rapidly drawn toward and delivered into the target cells, leading to efficient transfection.

PerkinElmer
For more information
877-PKI-NYSE
www.perkinelmer.com
http://science.labvelocity.com

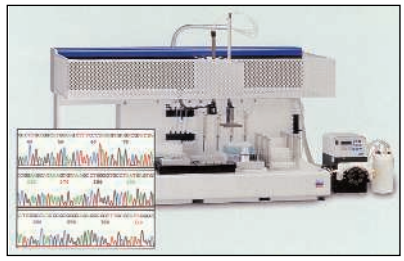
Protein Array Analysis Platform

The ProScanArray and ScanArray Gx series of microarray scanners support both proteomic- and genomic-array–based applications. The ProScanArray series of multi-application microarray scanners are state-of-the-art, upgradeable systems. These newly designed systems make use of proven confocal technology that collects more signal of interest so that background noise is automatically lowered, delivering a higher signal-to-noise ratio and better sensitivity. PerkinElmer has designed these microarray scanners to integrate image acquisition and analysis for proteomic and genomic applications through its software. For higher throughput applications, the ProScanArray HT features a patented 20-slide autoloader. The ScanArray Gx is a two-laser microarray scanner suitable for genomic applications. Optimized for traditional array applications, its software allows for fast and accurate spot-finding capabilities for improving the quality and accuracy of data.

Qiagen
For more information
800-426-8157
www.qiagen.com
http://science.labvelocity.com

Automated Plasma Minipreps

The BioRobot 3000 workstation and DirectPrep 96 BioRobot Kit provide a cost-effective, complete, automated solution for high-throughput plasmid minipreps in a 96-well format. The procedure enables reproducible purification of plasmid DNA that is ready for direct use in high-throughput sequencing applications. The DirectPrep 96 BioRobot Kit integrates novel lysis chemistry, eliminating the need for a lysate-clearing step, with an easily automated one-plate procedure. The integrated system enables cost-effective purification of sequencing-grade plasmid DNA from 96 samples per run. Typically, yields of up to 4 μg high-copy plasmid DNA can be obtained from 1.25 ml E. coli culture. The yield of plasmid DNA is sufficient for several sequencing reactions and additional archiving of the sample.



Hirata Corp
For more information
317-856-8600
www.hirata.com
http://science.labvelocity.com

Chemotaxis Research Device

EZ-TAXIScan is a desk-top, six-channel device to observe and film cell motion in real time for chemotaxis research. Aligned cells at the channel edge migrate over a glass plate under a silicon chip made with semiconductor technology and images are taken by a charge-coupled device camera. Assays can be performed with fewer than 100 cells, suitable for research involving scarce samples. Individual cells are automatically counted and data are formatted and uploaded into the personal computer system, from which summary charts can be produced instantly.

BD Biosciences
For more information
858-812-8800
www.bdbiosciences.com
http://science.labvelocity.com

Multiplex Bead Array System

The BD Cytometric Bead Array Flex Set system is an open and configurable bead-based immunoassay system designed to be an easy method of creating multiplex assays. Users can choose from a variety of beads to customize their multiplex assays.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and government organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier by visiting http://science.labvelocity.com on the Web, where you can request that the information be sent to you by e-mail, fax, mail, or telephone.

Published by AAAS